REV: INFORMATION-THEORETIC EVALUATION OF FREE-TEXT RATIONALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Free-text rationales are a promising step towards explainable AI, yet their evaluation remains an open research problem. While existing metrics have mostly focused on measuring the direct association between the rationale and a given label, we argue that an ideal metric should also be able to focus on the new information uniquely provided in the rationale that is otherwise not provided in the input or the label. We investigate this research problem from an information-theoretic perspective using the conditional $\mathcal{V}$-information (Hewitt et al., 2021). More concretely, we propose a metric called REV (Rationale Evaluation with conditional $\mathcal{V}$-information), that can quantify the new information in a rationale supporting a given label *beyond* the information already available in the input or the label. Experiments on reasoning tasks across four benchmarks, including few-shot prompting with GPT-3, demonstrate the effectiveness of REV in evaluating different types of rationale-label pairs, compared to existing metrics. Through several quantitative comparisons, we demonstrate the capability of REV in providing more sensitive measurements of new information in free-text rationales with respect to a label. Furthermore, REV is consistent with human judgments on rationale evaluations. Overall, when used alongside traditional performance metrics, REV provides deeper insights into a model's reasoning and prediction processes.

1 INTRODUCTION

Model explanations have been indispensable for trust and interpretability in AI (Lipton, 2018; Doshi-Velez & Kim, 2017; Kim et al., 2016; Alvarez Melis & Jaakkola, 2018). Free-text rationales, which explain a model prediction in natural language, have been especially appealing due to their flexibility in eliciting the reasoning process behind the model's decision making (Camburu et al., 2018; Narang et al., 2020; Rajani et al., 2019; Kumar & Talukdar, 2020b; Brahman et al., 2021), making them closer to human explanations. However, current automatic evaluation of free-text rationales remains narrowly focused. Existing metrics primarily measure the extent to which a rationale can help a (proxy) model predict the label it explains (i.e., accuracy based) (Hase et al., 2020; Wiegreffe et al., 2021). Yet, these metrics offer little understanding of the *new information* contained in the rationale, as added to the original input, that could *explain why the label is selected*—the very purpose a rationale is designed to serve. For instance, the two rationales r_1^* and $\hat{r}_{1,a}$ in Fig. 1 would be considered equally valuable under existing metrics, even though they supply different amount of novel and relevant information.

In this paper, we overcome this shortcoming by introducing an automatic evaluation for free-text rationales along two dimensions: (1) whether the rationale supports (i.e., is predictive of) the intended label, and (2) how much *new information* does it provide to justify the label, **beyond** what is contained in the input. For example, rationale $\hat{r}_{1,b}$ in Fig. 1 violates (1) because it is not predictive of the label, "enjoy nature". Rationale $\hat{r}_{1,a}$ does support the label but contains no new information that justifies it, *beyond* what is stated in the input x; thus, it violates (2). Rationale r_1^* is satisfied along both dimensions: it supports the label and does so by providing new and relevant information, beyond what is in the input. Our proposed evaluation is designed to penalize both $\hat{r}_{1,a}$ and $\hat{r}_{1,b}$, while rewarding rationales like r_1^*.

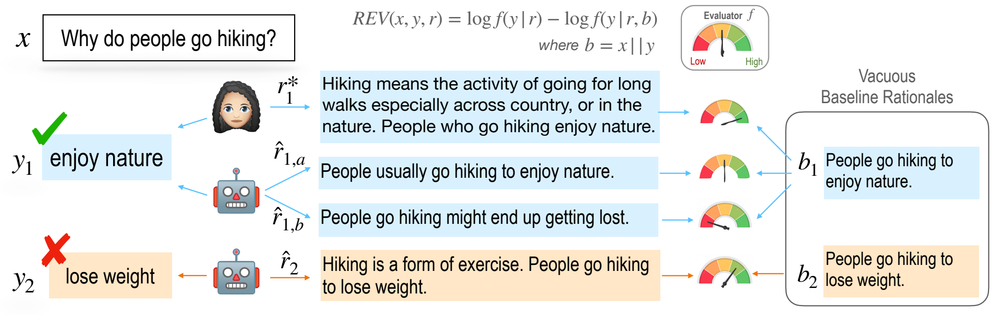

x
Why do people go hiking?
REV(x, y, r) = log f(y | r) − log f(y | r, b)
where b = x || y
Evaluator f
Low
High
Vacuous
Baseline Rationales
y1
enjoy nature
y2
lose weight
r*1
r̂1,a
r̂1,b
r̂2
Hiking means the activity of going for long walks especially across country, or in the nature. People who go hiking enjoy nature.
People usually go hiking to enjoy nature.
People go hiking might end up getting lost.
Hiking is a form of exercise. People go hiking to lose weight.
b1
People go hiking to enjoy nature.
b2
People go hiking to lose weight.

Figure 1: Our evaluation framework for different free-text rationales (r). r_1^* is a human-written rationale, $\hat{r}_{1,a}$ and $\hat{r}_{1,b}$ are two generated rationales for the true label y_1. Our metric, Rev, based on CVI (Hewitt et al., 2021) is able to distinguish all three rationales by measuring how much new and relevant information each adds over a vacuous rationale, b; performance-based evaluations can only distinguish between $\hat{r}_{1,a}$ and $\hat{r}_{1,b}$. For an (arguably) incorrect label, y_2, Rev still gives a positive score highlighting that $\hat{r}_2$ is able to provide new information for why it supports y_2. Prediction accuracy can be augmented with Rev to provide a fuller interpretability of model decisions.

We introduce a Rev[1], which adapts an information-theoretic framework from Xu et al. (2020) for evaluating free-text rationales along the two dimensions mentioned above. Specifically, Rev is based on conditional $\mathcal{V}$-information (Hewitt et al., 2021), which quantifies the degree of information contained in a representation, *beyond* another (baseline) representation, accessible to a model family, $\mathcal{V}$. As our baseline representation, we consider any vacuous rationale which simply combines an input with a given label, without providing any new information relevant to answering why the label was chosen. Rev adapts conditional $\mathcal{V}$-information to evaluate rationales, where the representation is obtained via an evaluator model trained to produce a label given the rationale. Other metrics do not take into consideration vacuous rationales, and are hence unable to measure new, label-relevant information in rationales, beyond a vacuous baseline.

Our experiments present evaluations with Rev for rationales under two reasoning tasks, commonsense question-answering (CQA; Talmor et al., 2018) and natural language inference (NLI; Bowman et al., 2015), across four benchmarks. Several quantitative evaluations demonstrate the capabilities of Rev in providing evaluations along new dimensions for free-text rationales, while being more consistent with human judgements compared to existing metrics. We also provide comparisons to demonstrate the sensitivity of Rev to various degrees of input perturbations. Additionally, our evaluation with Rev offers insights into why rationales obtained through chain-of-thought prompting (Wei et al., 2022) do not necessarily improve prediction performance. We will make our code and data public.

2 Rev: Information-Theoretic Evaluation of Rationales

We introduce a new metric, Rev, Rationale Evaluation with conditional $\mathcal{V}$-information, for evaluation of free-text rationales on the proposed dimensions (§2.2), based on the information-theoretic framework of conditional $\mathcal{V}$-information (§2.1).

We consider the setting where we have input $X \in \mathcal{X}$, label $Y \in \mathcal{Y}$, and free-text rationale $R \in \mathcal{R}$ generated for label Y. A common strategy to evaluate rationales R is through an evaluator $f \in \mathcal{V}$ based on how much R helps f predict Y given X. The evaluator $f : Z \rightarrow Y$ maps a variable Z to a label distribution. The definition of Z depends on the evaluation framework; e.g., Z can be a concatenation of X and R. The evaluator f is trained on a set of input, label and rationale triples $\mathcal{D}_{\text{train}} = \{(x_j, y_j, r_j)\}$, and applied to $\mathcal{D}_{\text{test}} = \{(x_i, y_i, r_i)\}$ for evaluation. The utility of R is formulated as the difference between the performance of the evaluator on predicting Y with R, and without it, i.e.

$$\text{Perf}[f(Y|X,R)] - \text{Perf}[f(Y|X)]. \tag{1}$$

[1] For Rationale Evaluation with conditional $\mathcal{V}$-information.

A larger performance gap indicates a better rationale. Existing metrics (Hase et al., 2020; Wiegreffe et al., 2021) compute the performance gap based on prediction accuracies, measuring how much R can help the evaluator correctly predict Y given X.

However, accuracy-based evaluation can only indicate whether or not a rationale is predictive of a label, but cannot quantify how much *new information the rationale provides to justify the label*. Fig. 1 illustrates this issue via an example. Accuracy-based evaluation can distinguish between $\hat{r}_{1,a}$ and $\hat{r}_{1,b}$ since $\hat{r}_{1,a}$ supports y_1 and $\hat{r}_{1,b}$ does not. However, it is unable to distinguish between r_1^* and $\hat{r}_{1,a}$ (since both are predictive of y_1), despite the fact that $\hat{r}_{1,a}$ does not provide any unique and relevant information to answer why the label should be y_1. In practice, vacuous rationales such as $\hat{r}_{1,a}$ are commonly seen in model generations (Sun et al., 2022; Wiegreffe & Marasović, 2021). This calls for an evaluation metric which is able to identify and penalize such vacuous rationales.

2.1 An Information-Theoretic Perspective on Rationale Evaluation

The key quantity of interest for our evaluation of rationales R is the amount of new information expressed in R (e.g., background knowledge, reasoning process) that can justify a label Y. The mutual information between R and Y, $I(Y;R)$ can be helpful for evaluating this quantity. However, we are not interested in the information that is already captured in the input X. A **vacuous** rationale, such as $\hat{r}_{1,a}$ in Fig. 1, which simply combines the input X and the label, Y, captures all the information in X and Y without specifying any new information to help understand why Y has been chosen for X; let us denote such rationales as $B \in \mathcal{B}$. Thus, we argue that a good evaluation metric must be able to measure the amount of relevant, new information contained in a rationale *beyond* what is contained in any vacuous rationale, B, that leads to the prediction of Y. Then the new information in R beyond what is available in B can be grounded with conditional mutual information (Shannon, 1948) as follows,

$$I(Y;R \mid B) = I(Y;R,B) - I(Y;B), \tag{2}$$

where the difference of two information quantities demonstrates the performance gap in Equation 1. Directly computing mutual information, however, is challenging because true distributions of random variables are usually unknown, and we do not have unbounded computation. A recently introduced information-theoretic framework called $\mathcal{V}$-information circumvents this by restricting the computation to certain predictive model families, $\mathcal{V}$ (Xu et al., 2020). Our approach to evaluate rationales extends this framework, following (Hewitt et al., 2021), as described below.

Conditional $\mathcal{V}$-information Given a model family $\mathcal{V}$ that maps two random variables R and Y, $\mathcal{V}$-information defines the usable information that can be extracted from R by models in $\mathcal{V}$ to predict Y, i.e. $I_{\mathcal{V}}(R \to Y)$. If $\mathcal{V}$ generalizes to the set of all possible functions, then $\mathcal{V}$-information is mutual information (Shannon, 1948). In practice, it is feasible to estimate the usable information from R about Y by selecting any neural model without frozen parameters as $\mathcal{V}$.[2]

Following conditional mutual information in information theory (Cover & Thomas, 2006), $\mathcal{V}$-information has been extended to conditional $\mathcal{V}$-information (CVI; Hewitt et al., 2021). CVI quantifies the $\mathcal{V}$-usable information in R about Y conditioned on a variable B, i.e.

$$I_{\mathcal{V}}(R \to Y \mid B) = H_{\mathcal{V}}(Y \mid B) - H_{\mathcal{V}}(Y \mid R, B). \tag{3}$$

Here B is any vacuous rationale that leads to the prediction of Y. In this work, we consider B simply as the concatenation of X and Y. We leave analyzing how different baseline construction impacts our metric to future work. $H_{\mathcal{V}}(\cdot \mid \cdot)$ is the conditional $\mathcal{V}$-entropy (Xu et al., 2020; Hewitt et al., 2021; Ethayarajh et al., 2022), defined as

$$H_{\mathcal{V}}(Y \mid B) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[b](y)]; \quad H_{\mathcal{V}}(Y \mid R, B) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[r,b](y)], \tag{4}$$

where $f[b]$ and $f[r,b]$ produce a probability distribution over the labels given b and $[r,b]$ as inputs respectively.[3] Further, we consider pointwise CVI for evaluating individual samples, (r,y,b) as

$$PI_{\mathcal{V}}(r \to y \mid b) = \log f[r,b](y) - \log f[b](y). \tag{5}$$

[2] Please see (Xu et al., 2020) for a detailed discussion of properties such as optional ignorance that a predictive family $\mathcal{V}$ must follow.

[3] Please see Appendix A for further details on CVI.

2.2 Computing Rev for Rationale Evaluation

Building on the framework of CVI, we propose a new metric REV, for Rationale Evaluation with conditional $\mathcal{V}$-information. We compute REV over a given test set, $\mathcal{D}_{\text{test}} = \{(x_i, y_i, r_i)\}$, by estimating CVI over the set with an evaluator $f \in \mathcal{V}$. For a test example (x, y, r), the REV score denoted as $\text{REV}(x, y, r)$ is computed based on Equation 5, where b is constructed by combining x and y.

$$\text{REV}(x, y, r) = PI_{\mathcal{V}}(r \rightarrow y \mid b) \tag{6}$$

The REV score for the test corpus $\mathcal{D}_{\text{test}}$, is given by the average pointwise REV score:

$$\text{REV} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_i \text{REV}(x_i, y_i, r_i). \tag{7}$$

Algorithm 1 shows the process of computing both pointwise and aggregate REV scores.

The higher the REV score, the more additional (*new* and *relevant*) information the rationale r contains to explain the label beyond the baseline rationale b. $\text{REV}(x, y, r)$ can take positive, negative, or zero values. When $\text{REV}(x, y, r)) > 0$, the rationale supplies additional information for supporting the label (e.g., r_1^* in Fig. 1); when $\text{REV}(x_i, y_i, r_i) = 0$, the rationale provides no additional information beyond the baseline (e.g., $r_{1,a}$ in Fig. 1); and when $\text{REV}(x_i, y_i, r_i) < 0$, the rationale contains additional information which does *not* support the label (e.g., $r_{1,b}$ in Fig. 1). REV can assign a positive score to a rationale for an incorrect prediction as long as the rationale supports it and provides additional information beyond a vacuous baseline rationale. Thus, REV cannot be seen as a replacement for prediction accuracy, but rather as an orthogonal metric to interpret the usefulness of a generated rationale for the model decision.

Constructing a Baseline with Vacuous Rationales Given an input x and a label y, we construct a baseline rationale b by converting x and y into a declarative sentence. For the CQA task, we adopt a pre-trained T5-3B model fine-tuned on a set of (*question*, *answer*, *declarative sentence*) tuples (Chen et al., 2021) [4] annotated by Demszky et al. (2018). For the NLI task, we use a template to convert (*premise*, *hypothesis*, *label*) tuple into a baseline rationale: "*premise* `implies` / `contradicts` / `is not related to` *hypothesis*". Table 1 shows some examples of constructed vacuous rationales.

Task	Input	Label	Baseline Rationale
CQA	Where can personal mushrooms be kept fresh?	refrigerator	Personal mushrooms can be kept fresh in the refrigerator.
NLI	Premise: A dog running in the surf. Hypothesise: A dog is at the beach.	entailment	A dog running in the surf implies a dog is at the beach.

Table 1: Examples of constructed baseline rationales for CQA and NLI tasks.

Training the evaluator, f We select a generative model $f \in \mathcal{V}$ as the evaluator to learn the mapping $f : [r, b] \rightarrow y$, where the input $[r, b]$ is the concatenation of r and b.[5] In particular, we use pre-trained language models (e.g., T5; Raffel et al., 2020) and fine-tune them on the training set $\mathcal{D}_{train} = \{(x, y^*, r^*)\}$, where $\{y^*\}$ and $\{r^*\}$ are gold labels and human-annotated rationales, respectively. We construct baseline rationales $\{b^*\}$ based on $\{(x, y^*)\}$. The objective is to maximize the log-likelihood of y^* given r^* and b^*.

After training, the evaluator can be applied to evaluate a given rationale-label pair (y, r) w.r.t. an input x. The rationale-label pair (y, r) can be model-generated and the label may not be ground-truth (e.g., y_2 in Fig. 1), while REV is still able to provide an assessment on the rationale along the two dimensions (§1), e.g., $\text{REV}(x, y_2, \hat{r}_2) = 0.6$ in Fig. 1.

[4] https://github.com/jifan-chen/QA-Verification-Via-NLI

[5] We do not train two models as Hewitt et al. (2021) did, one taking as input $[r, b]$ and the other taking as input b padded with dummy tokens. In our pilot experiments, the model trained solely with b can be overconfident on its predictions as b simply leaks the label information.

3 Experimental Setup

We outline our experimental setup by describing the reasoning tasks and datasets (§3.1), followed by the task and evaluation models (§3.2), and the baseline metrics for comparison (§3.3). Additional details on the setup are provided in Appendix B.

3.1 Datasets

We explore two reasoning tasks, namely CommonsenseQA (CQA) and Natural Language Inference (NLI) across four datasets, all containing human-annotated free-text rationales. For CQA task, we use ECQA (Aggarwal et al., 2021), CoS-E (v1.11; Rajani et al., 2019) and QuaRTz (Tafjord et al., 2019). For both ECQA and CoS-E, each commonsense question is paired with five candidate choices and the task is to select an answer from the candidates. ECQA contains higher quality human-written rationales compared to CoS-E (Aggarwal et al., 2021; Sun et al., 2022). QuaRTz is for open-domain reasoning about textual qualitative relationships, and the task is to select an answer from two options to the question based on the textual qualitative knowledge (rationale). e-SNLI (Camburu et al., 2018) provides explanations for the NLI, where given a premise, the task is to predict if a hypothesis entails, contradicts or is neutral to it. More details of the datasets are in Appendix B.1.

3.2 Task and Evaluation Models

Task models We choose T5 Large (Raffel et al., 2020) as the task model (finetuned on ground truth labels and rationales) to produce generated rationale-label pairs under three settings:

- XY*→R: Given an input text and the gold label, generate a rationale.
- X→YR: Given an input text, generate a label followed by a rationale. Since T5 decodes tokens sequentially, each R is generated conditioned on the predicted Y.
- X→RY: Given an input text, generate a rationale followed by a label. Here, we compute a likelihood for each candidate Y conditioned on R, and then select the most probable candidate. This operation can improve the model prediction accuracy, while weakening the consistency and relevance between the generated rationales and predicted labels.

After training, we collect three types of rationale-label pairs by applying the three task models on the test set of each dataset. In addition to these three settings, we also evaluate ground-truth labels paired with crowd-sourced rationales (Y*;R*).

Evaluators Our evaluator, f (see Equation 6 in §2) is also based on T5 Large trained on gold rationale-label pairs of the respective dataset. We refer readers to the Appendix C.1 for results using T5 Base, BART Large (Lewis et al., 2020) and GPT-2 Large (Radford et al., 2019) as the evaluator.

3.3 Other Metrics for Rationale Evaluation

We compare with two existing automatic metrics for free-text rationale evaluation: LAS (Hase et al., 2020) and RQ (Wiegreffe et al., 2021). Analogous to our evaluator, f, both approaches use proxy models; we use the same architecture (T5 Large) across metrics in our reported results.

Leakage-Adjusted Simulatability (LAS) Hase et al. (2020) evaluate the quality of free-text rationales via a proxy model, trained with the task model outputs as labels and original input texts combined with rationales as input sequences. The metric computes the difference between its prediction accuracy on the predicted label when the rationale is included into the input vs. when it is not, $\mathbb{1}[\hat{y} \mid x, \hat{r}] - \mathbb{1}[\hat{y} \mid x]$, averaged over examples grouped based on whether they leak labels or not. The final LAS score is given by the macro average across groups.

Rationale Quality (RQ) Wiegreffe et al. (2021) propose a variant of the simulatability in Hase et al. (2020). The main difference is that gold labels are used to train the model proxy and evaluate rationale quality. Specifically, the quality of a rationale $\hat{r}$ is measured as $\mathbb{1}[y^* \mid x, \hat{r}] - \mathbb{1}[y^* \mid x]$, where y^* is the gold label. Similarly, RQ is the average score over all test examples.

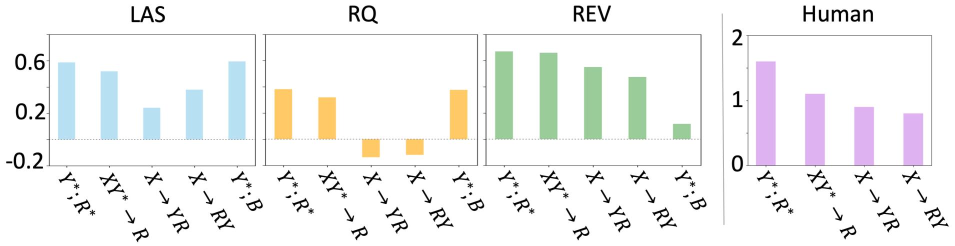


Figure 2: Left: automatic evaluation results of LAS, RQ and REV for rationale-label pairs on the ECQA test set. Right: human evaluation for rationale-label pairs on 230 randomly selected examples from the ECQA test set.

4 EXPERIMENTS

We first compare REV with existing metrics (§4.1) and human judgments (§4.2) on the ECQA dataset, as well as show REV on other CQA and NLI benchmarks. We then test the sensitivity of different metrics to input perturbations (§4.3). Next, we apply REV to generations via few-shot prompting (4.4). Additional experiments are listed in Appendix C.

4.1 COMPARISON BETWEEN EVALUATION METRICS

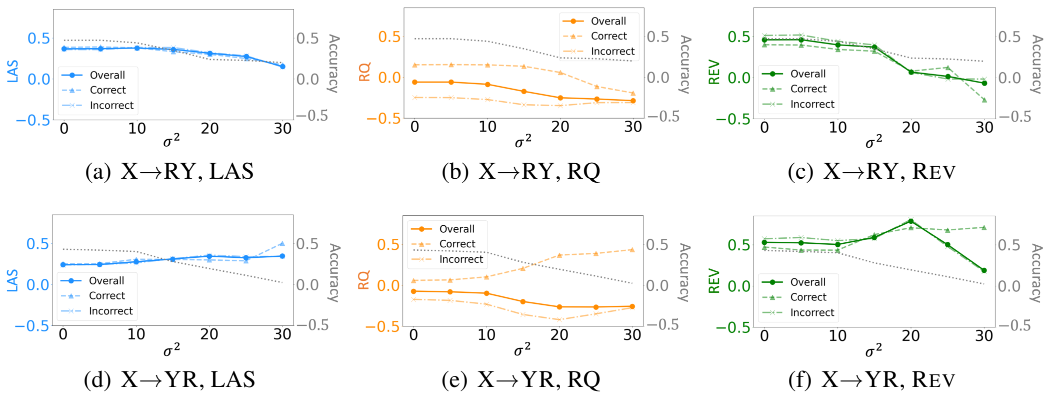
We compare REV to LAS and RQ, in evaluating different rationale-label pairs on the ECQA dataset. In addition to Y^*;R^*, $\mathrm{XY}^*\rightarrow\mathrm{R}$, X→YR and X→RY, we also explore the evaluation on vacuous baseline rationales (Y^*;B), which simply combine inputs and labels with no additional information. Note that the scores obtained from different metrics are not directly comparable due to different comparison scales and criteria (e.g., log-probability vs. accuracy). We mainly focus on the ranking over different types of rationale-label pairs. The results averaged over 4 random seeds are shown in the left part of Fig. 2. Qualitative results are provided in Table 5 in Appendix C.2.

All three metrics agree that the crowdsourced rationales (Y^*;R^*) in the ECQA have the highest quality. While by definition, REV for vacuous rationales is low, both LAS and RQ scores for these rationales are quite high, showing that these metrics are incapable of measuring the amount of additional information in rationales. Intuitively, we expect weaker rationale-label consistency in X→RY setting compared to X→YR, as the labels are forcefully selected among the candidates as opposed to being freely generated by the task model (§3.2). While REV is able to capture this intuition and rank X→YR higher than X→RY, LAS and RQ have a different ranking.

Datasets	Rationale-label pairs			
	Y^*;R^*	$\mathrm{XY}^*\rightarrow\mathrm{R}$	X→YR	X→RY
ECQA	0.6684	0.6401	0.5285	0.4586
CoS-E	0.3476	0.4576	0.3328	0.1518
QuaRTz	0.1851	0.1896	0.1624	0.1572
e-SNLI	1.1e-6	1.1e-6	1.08e-6	1.09e-6

Table 2: REV scores of different types of rationale-label pairs on the four datasets.

Next, we apply REV to evaluate crowdsourced and model generated rationale-label pairs (Y^*;R^*, $\mathrm{XY}^*\rightarrow\mathrm{R}$, X→YR, X→RY) across different datasets. For each dataset, the evaluator is trained on the training set with gold labels and crowdsourced rationales. The results are shown in Table 2. We observe that the gold rationales in the ECQA dataset achieve higher REV score than those in CoS-E. This observation is in line with the known quality issues of crowdsourced rationales in CoS-E (Aggarwal et al., 2021; Sun et al., 2022). Moreover, training the evaluator with CoS-E results in lower REV for all models, compared to training with ECQA. Interestingly, model-generated rationales ($\mathrm{XY}^*\rightarrow\mathrm{R}$) have higher REV scores than crowdsourced rationales for CoS-E (see examples in Table 6), and similar REV scores for ECQA, QuaRTz and e-SNLI. For QuaRTz, the model generated rationales seem to not contain much new and relevant information over a vacuous baseline. In the case of e-SNLI, the problem is even severer as most of the crowdsourced or generated rationales do not provide reasoning but rather follow a label-specific template e.g., *A implies (that) B* (Kumar & Talukdar, 2020a; Brahman et al., 2021).



(a) X→RY, LAS (b) X→RY, RQ (c) X→RY, REV

(d) X→YR, LAS (e) X→YR, RQ (f) X→YR, REV

Figure 3: Sensitivity test results of REV, LAS and RQ for different sources of rationale-label pairs on the ECQA dataset. The X-axis shows different levels of noise (σ^2). We plot the curve of Accuracy (model prediction accuracy) vs. Noise in gray dashed line. We also separate the evaluation results on populations on which the model predictions are correct ("Correct") or incorrect ("Incorrect") in addition to the overall evaluation on all test examples ("Overall").

4.2 HUMAN EVALUATION

To understand how REV correlates with human judgments of rationales, we conduct a crowdsourcing experiment via Amazon Mechanical Turk. We randomly sample 230 examples from the ECQA test set and ask workers to evaluate the four types of rationale-label pairs (Y*;R*, XY*→R, X→YR, X→RY) for each example. [6] We present workers with a question (input text), an answer (label) and an explanation (rationale), and ask them whether the explanation justifies the answer (*yes/no*). If they answer *yes*, we further ask them to evaluate the amount of additional information supplied by the explanation that explains *why* the answer might have been chosen for the question. The workers choose from *none* / *little* / *some* / *enough*, corresponding to a 4-point Likert-scale. We collect 3 annotations per instance and use majority vote to decide whether the rationale can justify the label. If *yes*, we take the average over the 3 human-annotated scores as the amount of information. Otherwise, we give a score of -1. More details of human evaluation are in Appendix C.6.

The results are shown in the right part of Fig. 2, where the ranking of the four types of rationale-label pairs is Y*;R* > XY*→R > X→YR > X→RY. While LAS and RQ rank X→RY better than X→YR (see the left part of Fig. 2), the ranking from REV is more consistent with human judgments, suggesting its effectiveness in evaluating rationale quality.

4.3 IS REV SENSITIVE TO INPUT PERTURBATIONS?

We test the sensitivity of all automatic metrics for rationale evaluation metrics to input (X) perturbations in the task model, under two settings: X→YR and X→RY. Following Wiegreffe et al. (2021), we add zero-mean Gaussian noise $\mathcal{N}(0, \sigma^2)$ to input word embeddings during inference, inducing task models to produce progressively degenerate rationales and labels. A good metric should be sensitive to the change of rationales and labels and reflect their relationships under input perturbations.

REV and RQ show similar trends as for X→RY in Fig. 3 (b) and (c). However, LAS is less sensitive to noise for both joint models, X→RY and X→YR, in Fig. 3 (a) and (d). Since the proxy model for LAS is trained on the task models' predicted labels and generated rationales, it can overfit to the degenerate rationale-label pairs under input perturbations, hence being less sensitive to input noise during inference.

The largest differences between REV and RQ are for X→YR. We observe the task model can predict incorrect labels and then make up reasonable-sounding rationales for its wrong predictions under certain input perturbations (e.g., when $\sigma^2 \le 20$); prior work also reports this finding (Narang et al.,

[6] We do not consider Y*;B because we have trained workers to recognize the baseline rationales as vacuous.

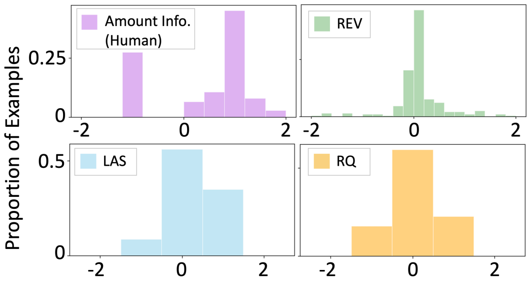


Figure 4: Histograms of human-annotated amount of information and pointwise Rev, LAS and RQ scores on the GPT-3 generated rationales for gold labels in few-shot prompting.

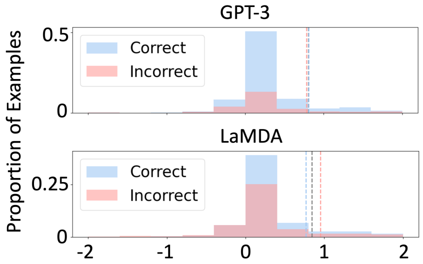


Figure 5: Distributions of Rev for rationales w.r.t. correct and incorrect predictions produced by GPT-3 and LaMDA respectively. Dashed lines indicate the average Rev for each group (blue/red) and the overall (gray).

2020; Wiegreffe et al., 2021). Rev does not drop under a certain amount of input perturbations (e.g., $\sigma^2 \leq 20$) in Fig. 3 (f), likely because the generated rationales still provide new information for describing both correct and incorrect labels (also see the example in Table 8). However, as the noise exceeds the certain level, Rev decreases indicating that the task model is perhaps no longer able to make up rationales for very noisy inputs. On the other hand, the behaviors of RQ and Rev are quite different in Fig. 3 (e) and (f). Since RQ is computed based on gold labels (§3.3), it has reduced sensitivity to input perturbations. When the prediction accuracy decreases, the overall evaluation of RQ is dominated by the results on incorrect predictions, as shown in Fig. 3 (e). We refer readers to the Table 8 in Appendix C.5 for qualitative analysis on sensitivity test.

4.4 Evaluating Rationales in Few-shot Prompting

We test the transferability of Rev in evaluating rationales generated by few-shot prompting, and get insights on the reasoning and prediction processes of large language models (e.g., GPT-3).

GPT-3 Rationales for Gold Labels. Wiegreffe et al. (2022) collected 250 high quality free-text rationales generated by few-shot prompting with GPT-3 (Brown et al., 2020) for CQA (given gold labels). Each example was assessed by 3 crowdworkers. We focus on two aspects of their annotations: "supports the gold label" and "amount of information". Crowdworkers provide a *yes / no* answer to justify whether a rationale supports the corresponding gold label. Only when the answer is *yes*, they are further asked to evaluate the amount of information contained in the rationale for justifying the label. The amount of information is roughly categorized into 3 levels: "Not Enough", "Enough", "Too Much", each annotated with a Likert-scale score.[7] In Fig. 4, we compare human annotation scores for amount of information[8] with the pointwise scores obtained by three automatic metrics, LAS, RQ, and Rev. For the automatic metrics, our evaluator, f and the proxy models of LAS and RQ are trained on the ECQA training set with gold labels and human-annotated rationales.

We observe that Rev provides finer-grained assessment of the information contained in rationales compared to LAS and RQ which only take {-1, 0, 1} values. When LAS and RQ are zero, it is unclear whether the rationale supports the label or not because the model proxy may predict the label based on the input only. The judgments of Rev on whether rationales support labels (Rev > 0) are close to human judgments (i.e., 69% vs. 73% support rate). The support rates of LAS and RQ are relatively low, i.e. 35% and 23%, while a large portion (56% and 60% respectively) corresponds to a zero LAS / RQ score.

Chain of Thought Rationales. Wei et al. (2022) propose *chain of thought prompting* to teach large language models to produce intermediate reasoning steps (rationales) before prediction, which

[7] The original human-annotated scores w.r.t. the three levels are: -1, 0, 1. Since Wiegreffe et al. (2022) suggest "a value of 0 is preferred to a value of 1", we map the scores {-1, 0, 1} to {0, 1, 2} accordingly. The value "-1" is then given to examples annotated as "not supporting gold labels".

[8] We take majority vote to decide "supports the gold label", and average the amount of information.

improves their prediction performance on a range of reasoning tasks (e.g., arithmetic and symbolic reasoning). However, the reported improvement is trivial for CQA (Wei et al., 2022), which motivates us to evaluate the intermediate rationales w.r.t. model predictions. We apply REV to analyze the generated rationales during intermediate reasoning steps and final predicted labels by GPT-3 text-davinci-002 (Brown et al., 2020) and LaMDA 137B (Thoppilan et al., 2022).[9]

Fig. 5 shows the distributions of REV for correctly and incorrectly predicted instances from GPT-3 and LaMDA, respectively. The prediction accuracy of GPT-3 is much higher than that of LaMDA (77% vs. 59%), while the average REV scores over all instances are close (0.80 vs. 0.84). For both GPT-3 and LaMDA, the REV distributions of correct and incorrect predictions are similar and most instances have positive REV scores. The results demonstrate the causality between the models' intermediate reasoning process and their final predictions, no matter whether the predicted labels are correct or incorrect. The average REV scores (dashed lines) over correct and incorrect predictions are close, especially for GPT-3. This is consistent with our observation that most generated rationales from the two models are describing their predicted labels. An insight we obtain is that the generated intermediate reasoning steps (rationales) support models' predictions (consistent REV scores), but cannot guarantee their correctness (discrepant accuracies). This partially explains the minor improvement of chain of thought prompting on CQA in Wei et al. (2022).

5 RELATED WORK

Model rationales broadly fall into two categories: extractive rationales and free-text rationales. Extractive rationales contain some important features extracted from input texts that make models produce final predictions (Lei et al., 2016; DeYoung et al., 2020; Jain et al., 2020; Schulz et al., 2019). Free-text rationales are produced by generative models in the form of natural language. Compared to extractive rationales, free-text rationales explain model predictions in a more human-like way and fill the gap in explaining reasoning tasks (Camburu et al., 2018; Narang et al., 2020; Rajani et al., 2019; Kumar & Talukdar, 2020b; Brahman et al., 2021).

Evaluations on extractive rationales have been well studied, generally from two perspectives — faithfulness and plausibility (DeYoung et al., 2020; Pruthi et al., 2022; Chan et al., 2022b). Faithfulness measures to which extent rationales reflect the true reasoning process of models, while plausibility evaluates how convincing rationales are to humans (Jacovi & Goldberg, 2020). Other perspectives include the ability of rationales in helping a student model simulate a teacher model (Pruthi et al., 2022) or bridging the communication between a classifier and a layperson (Treviso & Martins, 2020). Existing automatic metrics for free-text rationales focus on rationale-label association, and measure the utility of a rationale based on how much it helps a model proxy predict the given label (inspired by human simulatability Doshi-Velez & Kim (2017)) (Hase et al., 2020) or the gold label (Wiegreffe et al., 2021) given the input. Chan et al. (2022a) further propose a framework to evaluate the automatic metrics. However, none of them consider measuring the amount of additional information in free-text rationales. Sun et al. (2022) conduct a human study on the additional knowledge provided by free-text rationales. This work is the first that proposes an automatic metric to quantify the additional information in free-text rationales.

6 CONCLUSION

In this paper, we propose an information-theoretic metric, REV, to evaluate free-text rationale. REV measures if a rationale contains new information that is relevant for the label of interest, beyond the information contained in the input. We show the advantage of REV in evaluating different types of rationale-label pairs compared to existing metrics. We demonstrate that the evaluation of free-text rationales with REV is consistent with human judgments. REV also offers insights on evaluating rationales generated via few-shot prompting. In its current formulation, REV might reward a rationale for an incorrect prediction as long as the rationale supports the prediction with relevant additional information. Future work might explore evaluation that penalizes rationales which support incorrect predictions, thus bridging together predictive performance with interpretability metrics.

[9] Available at https://github.com/jasonwei20/chain-of-thought-prompting

REFERENCES

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. 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)*, pp. 3050–3065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL https://aclanthology.org/2021.acl-long.238.

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

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*, 2020.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi, and Xiang Ren. Frame: Evaluating simulatability metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*, 2022a.

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

Jifan Chen, Eunsol Choi, and Greg Durrett. Can nli models verify qa systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3841–3854, 2021.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley, 2nd edition, 2006.

Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020.

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

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning*, pp. 5988–6008. PMLR, 2022.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4351–4367, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.390. URL https://aclanthology.org/2020.findings-emnlp.390.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

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

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8730–8742, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771. URL https://aclanthology.org/2020.acl-main.771.

Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8730–8742, 2020b.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

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

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

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

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, 2019.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2019.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. Quartz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5941–5946, 2019.

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

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Marcos V Treviso and André FT Martins. The explanation game: Towards prediction explainability through sparse communication. *arXiv preprint arXiv:2004.13876*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Sarah Wiegreffe and Ana Marasović. Teach me to explain: A review of datasets for explainable natural language processing, 2021.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.47. URL https://aclanthology.org/2022.naacl-main.47.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

A Properties of Conditional $\mathcal{V}$-information

As proved by Hewitt et al. (2021), CVI has several useful properties:

1. *Non-Negativity*: $I_{\mathcal{V}}(R \to Y \mid B) \geq 0$.
2. *Independence*: If Y and B are jointly independent of R, then $I_{\mathcal{V}}(R \to Y \mid B) = 0$.
3. *Monotonicity*: If $\mathcal{U} \subseteq \mathcal{V}$, then $H_{\mathcal{V}}(Y \mid B) \leq H_{\mathcal{U}}(Y \mid B)$.

An implication from *Monotonicity* is complex models (e.g., pre-trained language models) can do better than simpler ones (e.g., linear models) in estimating $\mathcal{V}$-usable information. Since CVI measures the additional $\mathcal{V}$-usable information in R about Y beyond what's already extracted from B by models in $\mathcal{V}$, it grounds the goal of the proposed metric REV.

B Supplement of Experimental Setup

B.1 Datasets

For CQA task, we use ECQA (Aggarwal et al., 2021), CoS-E (v1.11) [10] (Rajani et al., 2019) and QuaRTz (Tafjord et al., 2019). Both ECQA and CoS-E originate from the CommonsenseQA dataset (Talmor et al., 2018), where each commonsense question is paired with 5 candidate choices and the task is to select an answer from the candidates. ECQA contains higher quality free-text rationales compared to CoS-E, in terms of comprehensiveness, coherence, non-redundancy, etc. (Aggarwal et al., 2021; Sun et al., 2022). QuaRTz is an open-domain reasoning task about textual qualitative relationships. Each instance contains a situated qualitative question, two answer options and a knowledge statement. The task is to select an answer from the two options to the question based on the textual qualitative knowledge. We use the knowledge statement as a free-text rationale since it explains why the answer is to the question. For NLI task, we use e-SNLI (Camburu et al., 2018) which is an extension of SNLI (Bowman et al., 2015) with augmented free-text human-written rationales. The task is to predict the entailment relationship between a premise and a hypothesis. Table 3 shows the summary statistics of the four datasets.[11]

Algorithm 1 Computing REV Scores

```
Input: evaluator $f$, test set $\mathcal{D}_{\text{test}} = \{(x_i, y_i, r_i)\}$
Initialize an empty list $\mathcal{S}$
for $(x_i, y_i, r_i) \in \mathcal{D}_{\text{test}}$ do
    Construct the baseline rationale $b_i$
    $\text{REV}(x_i, y_i, r_i)$
      $= \log f[r_i, b_i](y_i) - \log f[b_i](y_i)$
    $\mathcal{S}$.add($\text{REV}(x_i, y_i, r_i)$)
end for
$\text{REV} = \text{sum}(\mathcal{S}) / |\mathcal{S}|$
Output: $\mathcal{S}$, REV
```

Datasets	*#train*	*#dev*	*#test*
ECQA	7598	1090	2194
CoS-E	8766	975	1221
QuaRTz	2696	384	784
e-SNLI	54933	9842	9824

Table 3: Summary statistics of the datasets, where # counts the number of examples in the *train/dev/test* sets.

B.2 Models

We use Huggingface Transformers (Wolf et al., 2020) to access all task models and evaluators. We train each model for up to 30 epochs with a learning rate $5e-6$ and a batch size 8. All experiments

[10] We use the version v1.11 where each question is paired with 5 answer choices, for comparison with ECQA.

[11] Since CoS-E does not provide rationales for instances in the test set, we use the original development set as the test set and hold out 10% of training data as the new development set. We follow Hase et al. (2020) and randomly sample 10% of training data to form the training set for finetuning our models.

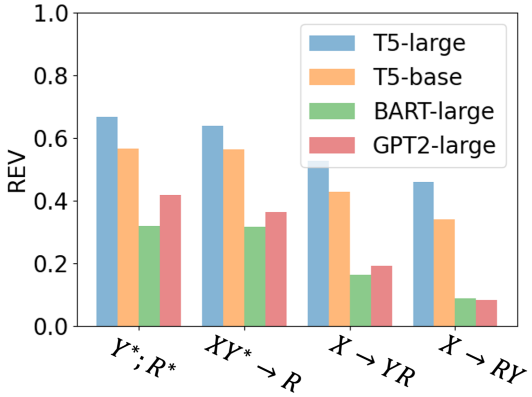


Figure 6: REV for evaluating rationale-label pairs on the ECQA dataset with different evaluator architectures.

were performed on a single NVIDIA RTX 8000 GPU. Table 4 shows input-output formattings of different task models for different tasks.

Type	Input	Output
XY*→R	CQA: [question] question [choice] choice-1 ... [choice] choice-n [answer] gold label [rationale] NLI: [premise] premise [hypothesis] hypothesis [answer] gold label [rationale]	rationale <eos>
X→YR	CQA: [question] question [choice] choice-1 ... [choice] choice-n [answer] NLI: [premise] premise [hypothesis] hypothesis [answer]	label [rationale] rationale <eos>
X→RY	CQA: [question] question [choice] choice-1 ... [choice] choice-n [rationale] NLI: [premise] premise [hypothesis] hypothesis [rationale]	rationale [answer] label <eos>

Table 4: The input-output formatting of different task models.

C Supplement of Experiments

C.1 Comparison Between Evaluator Architectures

We apply REV to evaluate different types of free-text rationales w.r.t. labels on the ECQA dataset. Figure 6 shows REV scores of the four types of rationale-label pairs evaluated by the four evaluators. The ranking of the four groups of rationale-label pairs is consistent across the four evaluators, i.e. Y*;R* > XY*→R > X→YR > X→RY. This ranking is also consistent with human evaluation in §4.2. Since ECQA contains high-quality crowdsourced rationales (Aggarwal et al., 2021), it is expected that the REV of gold rationale-label pairs (Y*;R*) is the highest. The REV of XY*→R is close to that of Y*;R*, indicating the task model (T5 Large) can produce good quality rationales when it is prompted with ground-truth labels. All four evaluators agree that the generated rationales of X→YR contain more additional background information for explaining the predicted labels than those of X→RY. This is consistent with our design of the X→RY in §3.3, where the generated rationales and labels have weakened relevance. For each type of rationale-label pairs, the four evaluators capture different amount of conditional $\mathcal{V}$-information, while T5 Large consistently outperforms other three models. In the reported experiments §4, we use T5 Large as the evaluator.

C.2 Qualitative Analysis of Different Metrics

Table 5 shows the qualitative analysis of different metrics on the four types of rationale-label pairs (Y*;R*, XY*→R, X→YR, X→RY) on the ECQA dataset. REV provides more accurate evaluations on those examples than LAS and RQ.

Type	Question	Label	Rationale	Metric		
				REV	LAS	RQ
Y*;R*	If you have a ticket and you are planning to eat hot dogs, where would you go?	baseball stadium	Hot dogs can be eaten at baseball stadium. When you go to a baseball stadium, you have a ticket and you may plan to eat hot dogs.	0.13	0	0
	How does a person go to space?	space shuttle	People go to space by a vehicle specially designed to travel to space. That vehicle is called a space shuttle.	0.01	0	0
	What is a dangerous outdoor activity for children?	sun themselves	Sunning themselves is a dangerous activity Children should not sun themselves	0.02	1	1
XY*→R	Where are old pictures kept?	attic	Attic is a place where old pictures are kept.	0.01	1	0
	What would you be if you comfort friend?	friendly	Comforting friend is a good thing.	-0.03	0	1
	What do customers do to a waiter after the waiter serves customers?	pay to	Paying to a waiter is the action of paying. Waiters get paid to serve customers.	1.59	-1	0
X→YR	Where is there likely to b more than one desk drawer?	desk	Desk drawer is a drawer used for storing office supplies. There is likely to be more than one desk drawer in office.	-0.34	-1	1
	What leads to someone's death when they are very depressed?	suicide	Suicide is the act of committing suicide. When someone is very depressed, suicide leads to their death.	0.25	0	0
	Where are you normally when you take a bath?	hotel room	Hotel room is a place where people stay. Bathing is normally done in hotel rooms.	0.01	0	-1
X→RY	What is likely heard by those going to a party?	laughter	People go to a party to meet new people. People are likely to hear laughter at the party.	0.49	1	0
	What would you do if you have excitement and do not want to stay in your house?	go to gym	Go to gym is to go to a place where you can express information. If you have excitement and do not want to stay in your house, then you would go somewhere.	0.21	1	0
	If you're caught committing murder, an injection can lead to your own what?	die	An injection can lead to one's own death. If you're caught committing murder, you can be injected into your own body and die.	0.29	0	0

Table 5: Pointwise evaluation of REV, LAS and RQ on different types of rationale-label pairs. Incorrect labels are colored red.

C.3 Qualitative Analysis of CoS-E Rationales

Table 6 shows the exemplar of REV scores for crowdsourced and model-generated (XY*→R) rationales for CoS-E. The main observation is model-generated rationales (XY*→R) generally support labels, though provide limited new information, while many crowdsourced rationales in CoS-E are noisy or uninformative. Specifically, compared to the crowdsourced rationales in CoS-E, we observe that XY*→R can produce better rationales that support the labels, which also corresponds to higher REV scores. However, the new information contained in those rationales is still limited (please see examples). A possible reason is the task model (XY*→R) hardly learns to produce more informative rationales when trained using lower quality rationales from CoS-E, known quality issue as reported in prior work (Aggarwal et al., 2021; Sun et al., 2022).

Type	Input	Label	Rationale	REV
Crowdsourced	The goal was to hit the target, but a projectile ball can't hit anything if it isn't in what?	motion	if you stand still you get hit	0.1
	When you get together with friends to watch film, you might do plenty of this?	have fun	when the working day is done	-0.4
	They dealt with combustible mixtures in their experiments, this is why they kept a fire extinguisher where?	chemistry lab	mixtures mixing fruitsa	-0.1
$XY^* \rightarrow R$	The goal was to hit the target, but a projectile ball can't hit anything if it isn't in what?	motion	a projectile ball can't hit anything if it's not in motion	0.5
	When you get together with friends to watch film, you might do plenty of this?	have fun	when you get together with friends to watch film, you might do plenty of fun	0.7
	They dealt with combustible mixtures in their experiments, this is why they kept a fire extinguisher where?	chemistry lab	chemistry labs deal with combustible mixtures in their experiments.	0.5

Table 6: Exemplar of REV scores for crowdsourced and model-generated (XY*→R) rationales for CoS-E.

C.4 Qualitative Analysis of Negative Rev Scores

Table 7 shows some examples of X→RY with negative Rev scores on the ECQA dataset. When Rev < 0, we observe in most cases the rationale does not support the given label, while indicating other labels, or something even beyond the label candidates (e.g., "helicopter" in the second example). Or they could repeat the input (e.g., the first example). The same observation holds for other types of rationale-label pairs.

Input	Label	Rationale	Rev
What do people call it when they are going for run?	falling down	People call it run when they are going for run.	-1.6
What enables most people to transport themselves?	own cars	People who believe in god are able to transport themselves through helicopter.	-1.6
Where would you expect to find popcorn in a public place?	movie theater	Popcorn can be found in a public place.	-1.5
What are you usually at when you sit on a bench on a curb?	city	Ohio is a state in the United States. You are usually at street corner when you sit on bench on curb.	-0.8

Table 7: Exemplar of negative REV scores for rationale-label pairs from X→RY on the ECQA dataset.

C.5 Qualitative Results of Sensitivity Test

Table 8 shows some examples from the sensitivity test in §4.3.

C.6 Human Evaluation Details

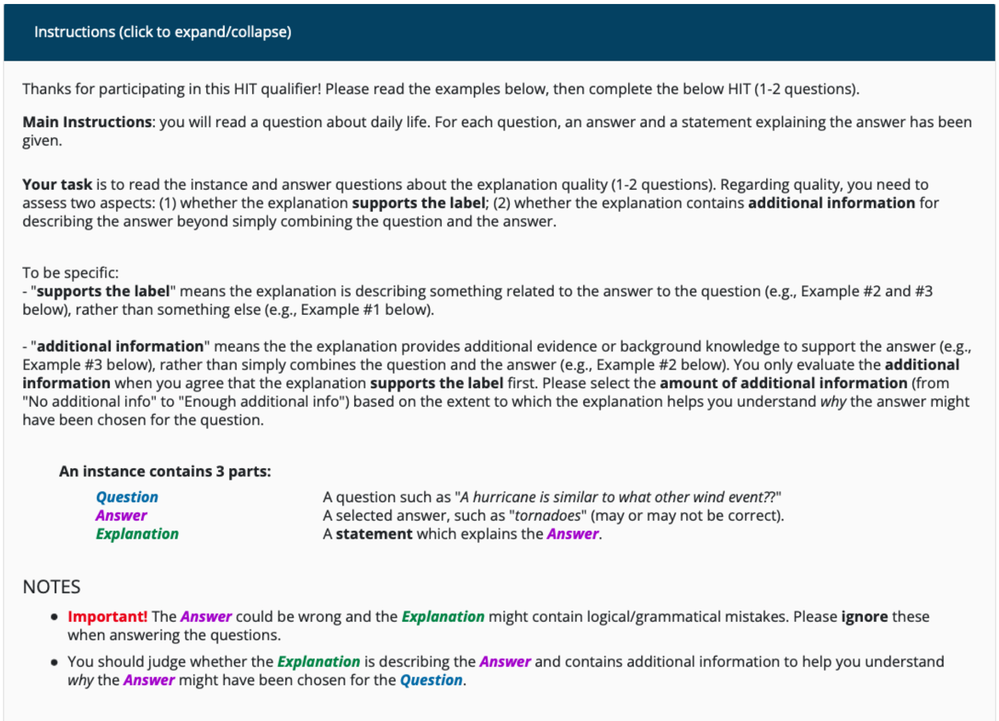
We randomly select 230 examples from the ECQA test set and conduct human evaluation on the four types of rationale-label pairs (Y^*;R^*, $XY^* \rightarrow R$, $X \rightarrow YR$, $X \rightarrow RY$) w.r.t. each example through the Amazon Mechanical Turk (AMT). Each instance is assessed by 3 workers. We pay the workers $0.08 for assessing each instance.

Figure 7 shows the instructions we provide to workers. In Figure 8, we show three examples, illustrating when the explanation (rationale) does not justify the answer (label), when the explanation supports the answer while not supplying additional information, and when the explanation supports the answer and provides additional information. Figure 9 shows the interface of the actual hit for human evaluation.

For each instance, we provide a question (input), an answer (label), and an explanation (rationale), and ask the workers to answer the following two questions:

1. *Does the Explanation justify the given Answer?* (yes or no) The question is to ask workers to judge whether the rationale supports the label or not.

2. *If yes, how much additional information does the Explanation have to justify the Answer beyond just reiterating what is stated in Question and Answer?* (No additional info, Little additional info, Some additional info, Enough additional info) We only ask this question if the workers choose "yes" for the first question. We design this question to ask workers to evaluate the extent to which the rationale provides additional information for justifying the label beyond repeating it w.r.t. the input.



Instructions (click to expand/collapse)

Thanks for participating in this HIT qualifier! Please read the examples below, then complete the below HIT (1-2 questions).

Main Instructions: you will read a question about daily life. For each question, an answer and a statement explaining the answer has been given.

Your task is to read the instance and answer questions about the explanation quality (1-2 questions). Regarding quality, you need to assess two aspects: (1) whether the explanation **supports the label**; (2) whether the explanation contains **additional information** for describing the answer beyond simply combining the question and the answer.

To be specific:
- "**supports the label**" means the explanation is describing something related to the answer to the question (e.g., Example #2 and #3 below), rather than something else (e.g., Example #1 below).

- "**additional information**" means the the explanation provides additional evidence or background knowledge to support the answer (e.g., Example #3 below), rather than simply combines the question and the answer (e.g., Example #2 below). You only evaluate the **additional information** when you agree that the explanation **supports the label** first. Please select the **amount of additional information** (from "No additional info" to "Enough additional info") based on the extent to which the explanation helps you understand *why* the answer might have been chosen for the question.

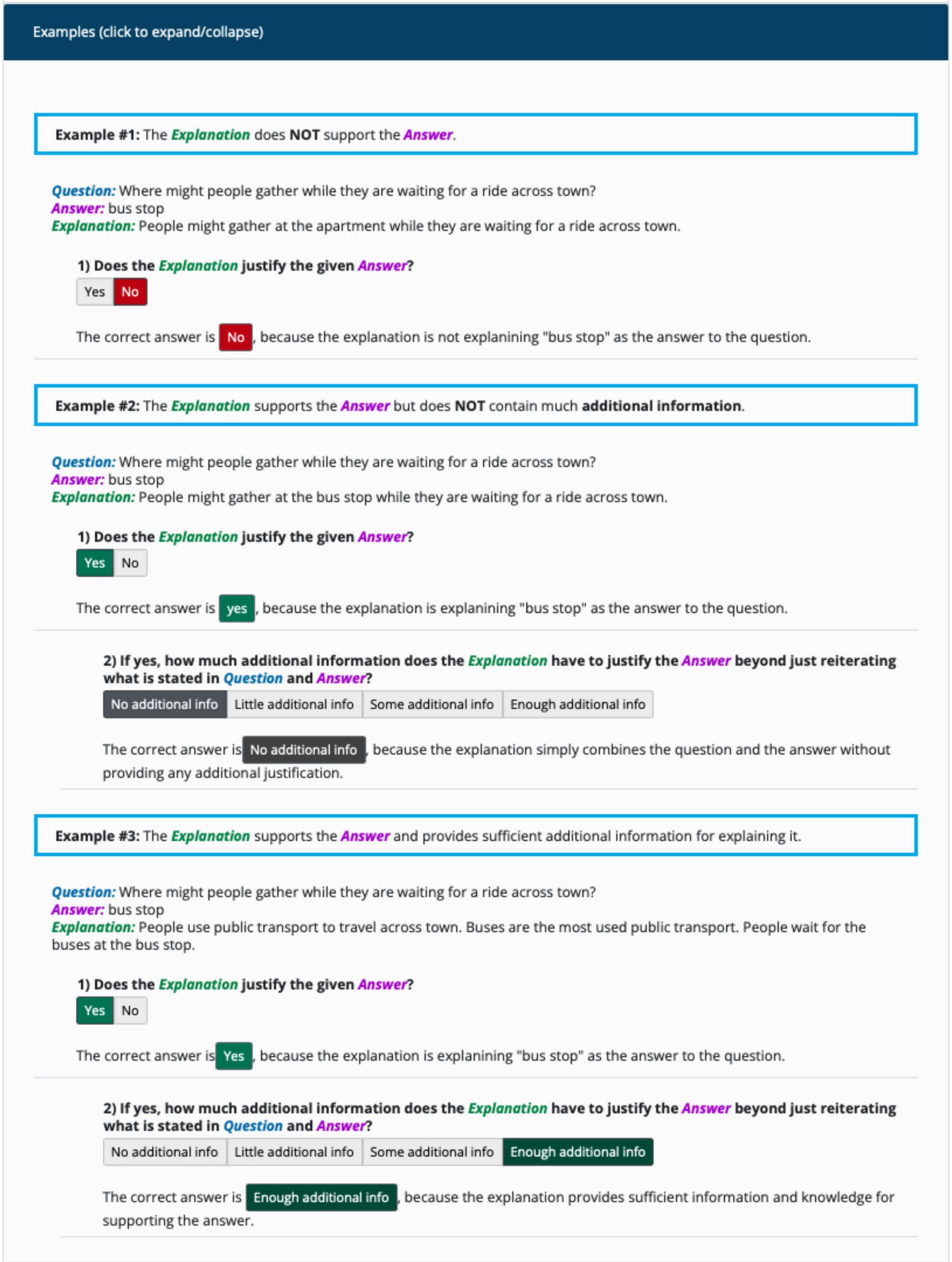
An instance contains 3 parts:

Question	A question such as "*A hurricane is similar to what other wind event??*"
Answer	A selected answer, such as "*tornadoes*" (may or may not be correct).
Explanation	A **statement** which explains the ***Answer***.

NOTES

- **Important!** The ***Answer*** could be wrong and the ***Explanation*** might contain logical/grammatical mistakes. Please **ignore** these when answering the questions.
- You should judge whether the ***Explanation*** is describing the ***Answer*** and contains additional information to help you understand *why* the ***Answer*** might have been chosen for the ***Question***.

Figure 7: The instructions of human evaluation in the user interface on AMT.



Examples (click to expand/collapse)

Example #1: The ***Explanation*** does **NOT** support the ***Answer***.

Question: Where might people gather while they are waiting for a ride across town?
Answer: bus stop
Explanation: People might gather at the apartment while they are waiting for a ride across town.

1) Does the *Explanation* justify the given *Answer*?

Yes [✓ No]

The correct answer is No, because the explanation is not explaining "bus stop" as the answer to the question.

Example #2: The ***Explanation*** supports the ***Answer*** but does **NOT** contain much **additional information**.

Question: Where might people gather while they are waiting for a ride across town?
Answer: bus stop
Explanation: People might gather at the bus stop while they are waiting for a ride across town.

1) Does the *Explanation* justify the given *Answer*?

[✓ Yes] No

The correct answer is yes, because the explanation is explanining "bus stop" as the answer to the question.

2) If yes, how much additional information does the *Explanation* have to justify the *Answer* beyond just reiterating what is stated in *Question* and *Answer*?

[✓ No additional info] Little additional info Some additional info Enough additional info

The correct answer is No additional info, because the explanation simply combines the question and the answer without providing any additional justification.

Example #3: The ***Explanation*** supports the ***Answer*** and provides sufficient additional information for explaining it.

Question: Where might people gather while they are waiting for a ride across town?
Answer: bus stop
Explanation: People use public transport to travel across town. Buses are the most used public transport. People wait for the buses at the bus stop.

1) Does the *Explanation* justify the given *Answer*?

[✓ Yes] No

The correct answer is Yes, because the explanation is explanining "bus stop" as the answer to the question.

2) If yes, how much additional information does the *Explanation* have to justify the *Answer* beyond just reiterating what is stated in *Question* and *Answer*?

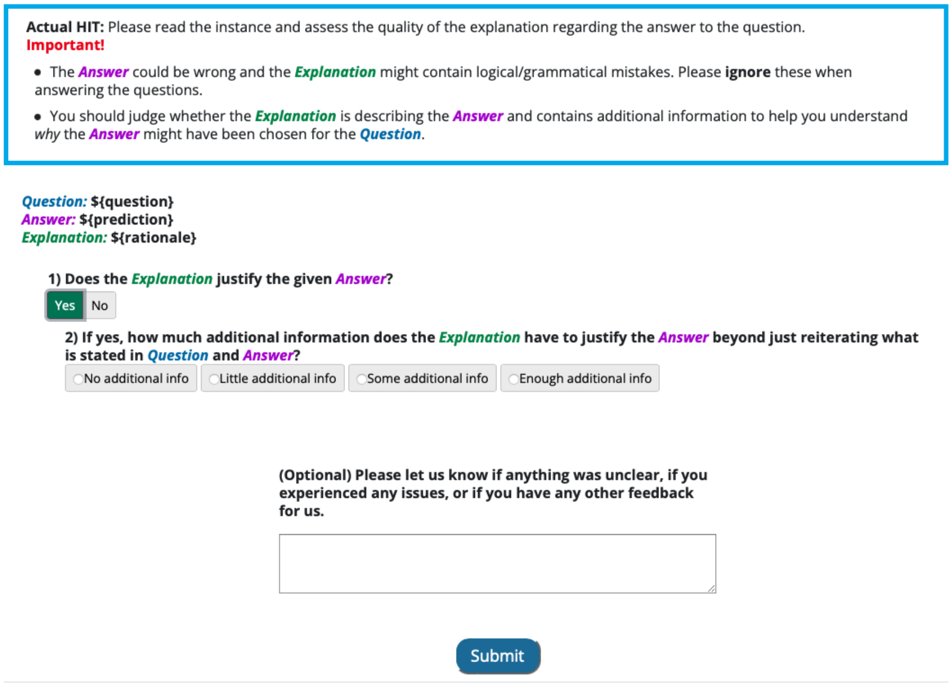
No additional info Little additional info Some additional info [✓ Enough additional info]

The correct answer is Enough additional info, because the explanation provides sufficient information and knowledge for supporting the answer.

Figure 8: Exemplars provided to worker in the user interface on AMT.

C.7 Additional Analysis on Label-Related Sentences

In some cases, a rationale contains the given label and provides new information related to the label, but does not necessarily explain why the label is selected for the input. To evaluate such rationales, we randomly select 250 gold labels in ECQA and extract their related sentences from a large-scale knowledge base—GenericsKB (Bhakthavatsalam et al., 2020). Those sentences contain the labels, while providing little or irrelevant new information to explain the labels w.r.t. the inputs. We use them as trivial rationales for evaluation. The average REV scores for those trivial rationales and their crowdsourced counterparts are 0.02 and 0.87 respectively, indicating the effectiveness of REV



Actual HIT: Please read the instance and assess the quality of the explanation regarding the answer to the question.
Important!

- The ***Answer*** could be wrong and the ***Explanation*** might contain logical/grammatical mistakes. Please **ignore** these when answering the questions.
- You should judge whether the ***Explanation*** is describing the ***Answer*** and contains additional information to help you understand *why* the ***Answer*** might have been chosen for the ***Question***.

Question: **${question}**
Answer: **${prediction}**
Explanation: **${rationale}**

1) Does the *Explanation* justify the given *Answer*?

Yes No

2) If yes, how much additional information does the *Explanation* have to justify the *Answer* beyond just reiterating what is stated in *Question* and *Answer*?

○ No additional info ○ Little additional info ○ Some additional info ○ Enough additional info

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any other feedback for us.

Figure 9: The actual hit of human evaluation in the user interface on AMT.

in identifying the new and relevant information in rationales. Table 9 shows the REV scores of some examples and the corresponding crowdsourced rationales. The results show that REV can distinguish the new information in different rationales and penalize meaningless rationales. Overall, REV gives higher scores to crowdsourced rationales than trivial sentences from GenericsKB.

X→YR
Question: What is likely to happen if you're doing housework all day?
Choices: boredom, tiredness, get tired, backache, get a clean house
Gold label: get tired
Gold rationale: Get tired means no longer wanting someone or wanting to do something because you are bored with it or annoyed by it. Get tired is likely to happen if you're doing housework all day.

$\sigma^2 = 0, \textbf{Rev} = 0.85$
Output: get tired [rationale] If you're doing housework all day, it's likely to happen if you are getting tired.

$\sigma^2 = 5, \textbf{Rev} = 0.85$
Output: get tired [rationale] If you're doing housework all day, it's likely to happen if you are getting tired.

$\sigma^2 = 10, \textbf{Rev} = 0.85$
Output: get tired [rationale] If you're doing housework all day, it is likely to happen if you are getting tired.

$\sigma^2 = 15, \textbf{Rev} = 0.30$
Output: backache [rationale] Backache is a condition of being tired. You are likely to get tired if you're doing housework all day.

$\sigma^2 = 20, \textbf{Rev} = 0.81$
Output: boredom [rationale] Boredom is a state of being or being unable to do something. People tend to get bored easily when they are bored.

$\sigma^2 = 25, \textbf{Rev} = 0.31$
Output: backache [rationale] Backache is the state of being unable to move backwards or forwards in a straight line. One cannot move forward if they are not able to do so.

$\sigma^2 = 30, \textbf{Rev} = 0$
Output: banishing oneself from one's own body by removing it from the body of another person. This is a way of preventing one from being harmed by other people. It is possible to get rid of people from their own bodies if they are living in the same house. That is why they have to live together.

X→RY
Question: What happens when not paying attention when bathing?
Choices: dry skin, using soap, reading newspaper, shampoo, slip and fall
Gold label: slip and fall
Gold rationale: Bathing is generally done in a bathroom. Bathrooms could be slippery due to water. If not careful slipperiness can cause one to slip and fall.

$\sigma^2 = 0, \textbf{Rev} = 0.90$
Output: Bathing can cause slip and fall when one is not paying attention. Slip and Fall is a common occurrence when people do not pay attention when bathing. [answer] slip and fall

$\sigma^2 = 5, \textbf{Rev} = 0.90$
Output: Bathing can cause slip and fall when one is not paying attention. Slip and Fall is a common occurrence when people do not pay attention when bathing. [answer] slip and fall

$\sigma^2 = 10, \textbf{Rev} = 0.90$
Output: Bathing can cause slip and fall when not paying attention. Slipping and falling is a result of not being attentive when bathing. [answer] slip and fall

$\sigma^2 = 15, \textbf{Rev} = 0.86$
Output: Bathing is a process of washing the body. Not paying attention means not paying enough attention to the bathing process. [answer] slip and fall

$\sigma^2 = 20, \textbf{Rev} = -0.45$
Output: Bathing is a process of washing and rinsing. Paying attention means paying attention to something. [answer] reading newspaper

$\sigma^2 = 25, \textbf{Rev} = -0.37$
Output: Paying attention to what you are doing is a sign of being aware of what is going on around you. Knowing what to do is to pay attention. [answer] dry skin

$\sigma^2 = 30, \textbf{Rev} = -1.11$
Output: Paying attention to one's own needs is a basic human need. One needs to be aware of the needs of others in order to meet them...engavengendersengendoenv -end-engagingendentenant [answer] slip and fall

Table 8: Examples of the two task models' outputs under different levels of noise in sensitivity test.

Source	Input	Label	Rationale	REV
Crowdsourced	There is normally a parking area near what collection of small dwellings?	apartment complex	Apartment complex is a small dwelling area. People park their vehicles in the complex's designated parking area at night.	0.3
	Where will you find billions of grains of sand?	beach	A beach is a pebbly or sandy shore, especially by the sea between high and low water marks. A sandy shore contains billions of grains of sand.	0.1
	What is a person the leader of after they are inagurated?	country	Country is a nation with its own government, occupying a particular territory. The person is the leader of a country after they are inaugurated.	0.6
GenericsKB	There is normally a parking area near what collection of small dwellings?	apartment complex	Apartment complexes are usually on easy-access roads and places to shop.	0.1
	Where will you find billions of grains of sand?	beach	Beach meshing operations change the habitat of some coastal areas, increasing mortality of sharks.	0
	What is a person the leader of after they are inagurated?	country	Countries adopt a peg as a way of promoting international confidence in their own currency.	-0.2

Table 9: Exemplar of REV scores for crowdsourced rationales and label-related sentences from GenericsKB for ECQA.