What to Learn, and How: Toward Effective Learning from Rationales

Samuel Carton
University of Chicago
carton@uchicago.edu

Surya Kanoria
University of Colorado Boulder
surya.kanoria@colorado.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Abstract

Learning from rationales seeks to augment model prediction accuracy using human-annotated rationales (i.e. subsets of input tokens) that justify their chosen labels, often in the form of intermediate or multitask supervision. While intuitive, this idea has proven elusive in practice. We make two observations about human rationales via empirical analyses: 1) maximizing rationale supervision accuracy is not necessarily the optimal objective for improving model accuracy; 2) human rationales vary in whether they provide sufficient information for the model to exploit for prediction. Building on these insights, we propose several novel loss functions and learning strategies, and evaluate their effectiveness on three datasets with human rationales. Our results demonstrate consistent improvements over baselines in both label and rationale accuracy, including a 3% accuracy improvement on MultiRC. Our work highlights the importance of understanding properties of human explanations and exploiting them accordingly in model training.

1 Introduction

In the past several years, explainability has become a prominent issue in machine learning, addressing concerns about the safety and ethics of using large, opaque models for decision-making. As interest has grown in explanations for understanding model behavior, so has interest grown in soliciting gold-standard explanations from human annotators and using them to inject useful inductive biases into models (Hase and Bansal, 2021). Many such explanation datasets have become available recently (Wiegreffe and Marasović, 2021).

A common format for explanations in NLP is the *rationale*, a subset of input tokens that are relevant to the decision. A popular architecture for generating such explanations is the *rationale model*, an explain-then-predict architecture which first extracts a rationale from the input and then makes a prediction from the rationale-masked text (that is, only the tokens included in rationale) (Lei et al., 2016; DeYoung et al., 2019). Without external supervision on this rationale, we typically pursue parsimony via a sparsity objective. Table 1A shows an example unsupervised rationale.

(A) Unsupervised rationale

[CLS] susan wanted to have a birthday party . she called all of her friends . she has five friends . her mom said that susan can invite them all to the party . her first friend could not go to the party because she was sick . her second friend was going out of town . her third friend was not so sure if her parents would let her . the fourth friend said maybe . the fifth friend could go to the party for sure . susan was a little sad . on the day of the party , all five friends showed up . each friend had a present for susan . susan was happy and sent each friend a thank you card the next week . [SEP] how many people did susan call ? | | 5 [SEP]

Prediction: False

(B) Human rationale

[CLS] susan wanted to have a birthday party . she called all of her friends . she has five friends . her mom said that susan can invite them all to the party . her first friend could not go to the party because she was sick . her second friend was going out of town . her third friend was not so sure if her parents would let her . the fourth friend said maybe . the fifth friend could go to the party for sure . susan was a little sad . on the day of the party , all five friends showed up . each friend had a present for susan . susan was happy and sent each friend a thank you card the next week . [SEP] how many people did susan call ? | | 5 [SEP]

Prediction: True

Table 1: An example of unsupervised versus human-provided rationale in MultiRC. The unsupervised model struggles to localize its attention and makes an incorrect prediction. The same model makes a correct prediction by only looking at the human rationale.

With the benefit of a human-annotated rationale for the true label, we can begin to understand model mistakes in terms of reliance on inappropriate features (and correct them). In the example above, the unsupervised rationale suggests that the model's

Findings of the Association for Computational Linguistics: ACL 2022, pages 1075 - 1088

mistake is due to missing key information about how many friends Susan has (i.e., "five"). Forcing the model to see these key tokens by only using the human rationale as the input fixes this mistake (Table 1B). Prior work has shown that this is not a fluke. For some datasets, human rationales consistently improve model accuracy over baseline when used as an input mask, by orienting model attention toward informative tokens and away from confounding ones (Carton et al., 2020).

Knowing that human rationales contain useful predictive signal, the key question becomes: **can we improve model prediction accuracy by incorporating human rationales into training?**

Numerous approaches to using human rationales in training have been tried, including: regularizing the parameters of a (linear) model (Zaidan et al., 2007); regularizing model output gradients (Ross et al., 2017); regularizing internal transformer attention weights (Jayaram and Allaway, 2021); and direct supervision on a rationale model (DeYoung et al., 2019), which serves as our baseline approach in this paper. These approaches have generally failed to significantly improve model prediction accuracy (Hase and Bansal, 2021).

A quality these prior approaches have in common is treating human rationales as *internally and collectively uniform* in predictive utility. That is, any token included in the human rationale is treated as equally important to include in the input representation; vice versa for tokens excluded. Furthermore, all human rationales are weighted equally.

The reality, we demonstrate empirically via ablation studies in §4, is that the predictive utility of human rationales is distributed unevenly between tokens in a rationale, and unevenly between rationales in a dataset. Based on this analysis, we suggest that learning objectives which weight every token equally (accuracy in the case of direct supervision), and every rationale equally, are not optimal for improving downstream model accuracy.

We operationalize these hypotheses in four distinct modifications to the baseline rationale model architecture. Three of these modify the naive token-wise accuracy supervision objective, and the fourth implements "selective supervision", ignoring unhelpful human rationales in training.

Evaluating on three datasets, our proposed methods produce varying levels of improvement over both a baseline BERT model and a baseline BERT-to-BERT supervised rationale model, ranging from substantial for MultiRC (3%) to marginal for E-SNLI (0.4%). Additionally, our methods also improve rationale prediction performance.

Taken together, our results demonstrate the importance of considering the variance of predictive utility both between and within human rationales as a source of additional training signal. Our proposed modifications help pave the way toward truly effective and general learning from rationales.

2 Related Work

2.1 Rationalization

The extractor-predictor rationale model proposed by Lei et al. (2016) and described in more detail in §5, is an approach to feature attribution, which is one among many families of explanation methods (see Vilone and Longo (2020) for a recent survey).

Recent work has extended the original architecture in various ways, including replacing the use of reinforcement learning with differentiable binary variables (Bastings et al., 2020; DeYoung et al., 2019), alternatives to the original sparsity objective (Paranjape et al., 2020; Antognini and Faltings, 2021), and additional modules which change the interaction dynamics between the extractor and predictor (Carton et al., 2018; Yu et al., 2019; Chang et al., 2020). Pipeline models (Lehman et al., 2019) are similar, but train the two modules separately rather than end-to-end.

Rationale models are a powerful approach to NLP explanations because of how specific objectives can be put on the properties of the rationale, but they have some downsides. First, they are unstable, the extractor often collapsing to all-0 or all-1 output (DeYoung et al., 2019; Yu et al., 2019). We introduce an engineering trick in §5 that appears to lessen this risk. Also, with end-to-end training comes the risk of information leakage between the extractor and predictor (Jethani et al., 2021; Hase et al., 2020; Yu et al., 2021). This idea of leakage plays a part in how we estimate explanation predictive utility in section §4.

2.2 Learning from Explanations

Wiegreffe and Marasović (2021) present a review of explainable NLP datasets, a number of which have been incorporated into the ERASER collection and benchmark (DeYoung et al., 2019).

Early work in learning from human explanations include Zaidan et al. (2007) and Druck et al. (2009), and a line of work termed "explanatory debugging"

(Kulesza et al., 2015; Lertvittayakumjorn and Toni, 2021). More recent work spans a variety of approaches, categorized by Hase and Bansal (2021) into regularization (e.g., Ross et al. (2017)), data augmentation (e.g., Hancock et al. (2018)), and supervision over intermediate outputs (e.g., DeYoung et al. (2019); Jayaram and Allaway (2021)).

Significant improvements to model accuracy as a result of explanation learning have proven elusive. Studies occasionally claim such improvement, such as Rieger et al. (2020), which observes general improvements on a medical vision task. More commonly their claims pertain to secondary objective such as explanation quality (e.g., Plumb et al. (2020)), robustness (e.g., Ross et al. (2017), Srivastava et al. (2020)), or few-shot learning (e.g., Yao et al. (2021)). Hase and Bansal (2021) gives an overview of the problem and discusses circumstances under which learning from explanations is liable to work. Our paper contributes to this discussion by considering the variance of training signal quality both within and between human rationales, and how to exploit these variances.

3 Data

We consider three datasets in this work. All three are document-query text comprehension tasks, where the task is to determine whether the query is true or false given the document. We use the train, development, test splits offered by DeYoung et al. (2019). Table 2 shows the basic statistics of each dataset based on the training set.

- **MultiRC** (Khashabi et al., 2018). A reading comprehension dataset of 32,091 document-question-answer triplets that are true or false. Rationales consist of 2-4 sentences from a document that are required to answer the given question.
- **FEVER** (Thorne et al., 2018). A fact verification dataset of 76,051 snippets of Wikipedia articles paired with claims that they support or refute. Rationales consist of a single contiguous subsnippet, so the basic unit of rationale is sentence.
- **E-SNLI** (Camburu et al., 2018). A textual entailment dataset of 568,939 short snippets and claims for which each snippet either refutes, supports, or is neutral toward. Input texts are much shorter than MultiRC and FEVER, and rationales are at the token level.

Dataset	Text length	Rationale length	Rationale granularity
MultiRC	336.0	52.0	sentence
FEVER	355.9	47.0	sentence
E-SNLI	23.5	6.1	token

Table 2: Basic statistics of the datasets.

4 Analysis

To understand properties of human rationales for the purpose of learning from rationales, we analyze the effect of human rationales when they are used as inputs to a trained model.

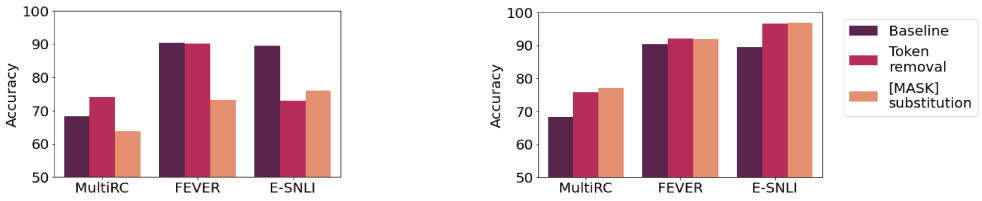
4.1 Human Rationales have Predictive Utility

A basic question about the viability of learning from rationales is whether human rationales bear the potential for improving model performance. That is, do human explanations successfully reveal useful tokens while occluding confounding tokens, such that a model evaluated only on the revealed tokens is able to get improved performance relative to the full input? We refer to such rationale-redacted inputs as *rationalized inputs*.

We define *sufficiency-accuracy* (**SA**) as how accurate the model is across a corpus of rationalized input. This is an aggregate measure, similar to *sufficiency* as defined in DeYoung et al. (2019) but focused on absolute performance rather than similarity to baseline model output. We refer to the sufficiency-accuracy of the human rationales as *human sufficiency-accuracy* (**HSA**).

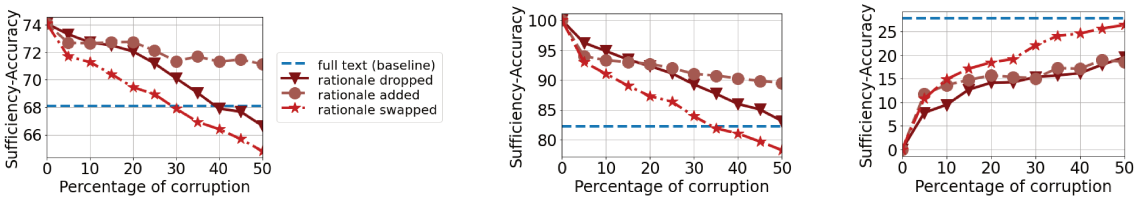
Estimating sufficiency-accuracy is problematic. The natural way to probe whether the tokens in a rationale are sufficient for an accurate prediction is to remove the non-included tokens from the input, run the model on just the included tokens, and assess its accuracy. But a version of the input where a majority of tokens are removed or masked (by a [MASK] special token in the case of BERT), is out-of-distribution relative to the training data, which has no removal or masking. This difference may lead to unpredictable output from the model when tested on masked input. This **masking-is-OOD** problem has not received much discussion in the literature, though Jacovi and Goldberg (2021) propose to mitigate it with random masking during model training. The effect of this problem will be to underestimate the sufficiency-accuracy of rationales tested against an un-adapted model.

The opposite problem stems from overfitting rather than OOD issues: **label leakage**. A human rationale may contain signal about the true label



(a) Fine-tuned on full input (unadapted). (b) Fine-tuned on both full and human-rationalized input (adapted).

Figure 1: Baseline performance vs. human sufficiency-accuracy for rationalized inputs with token removal and [MASK] token substitution. As rationalized inputs are different from the full text inputs that the original training set includes, we build a calibrated model where the model is trained on both full text inputs and rationalized inputs.



(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 2: Sufficiency-accuracy of human rationales on baseline BERT model with increasing levels of corruption via swaps, drops and additions. Model performance decreases quickly when we drop rationale tokens , but stays high as we add non-rationale tokens. These effects are moderated by HSA.

that goes beyond the semantics of the tokens included in the rationale, and a model trained on human-rationalized input may learn to pick up on these spurious signals. A known example is in E-SNLI, where annotators had different explanation instructions based on their chosen label. This issue is discussed in several recent papers (Yu et al., 2021; Jethani et al., 2021; Hase et al., 2020), albeit mostly concerning model-generated rather than human explanations. The effect of this problem will be to overestimate the sufficiency-accuracy of rationales tested against an adapted model.

Fig. 1 shows sufficiency-accuracy results for human rationales on both unadapted and adapted models. We expand on the analysis presented by Carton et al. (2020) by showing results for both masking-via-removal and masking-via-[MASK]-token-substitution.

Fig. 1a shows that token removal suffers less from the masking-is-OOD problem on an unadapted model than [MASK] token substitution. [MASK] token substitution results in lower accuracy across the board, while removal improves baseline accuracy for MultiRC, matches it for FEVER, and lowers it for E-SNLI.

With adaptation (Fig. 1b), token removal and [MASK] token substitution have near-identical effects, improving accuracy by a large margin for MultiRC and E-SNLI, and a small margin for FEVER. The near-100% sufficiency-accuracy for E-SNLI is probably due to label leakage.

If an unadapted model is liable to underestimate sufficiency model, and an adapted model to overestimate, then we suggest that the potential benefit of learning from rationales lies somewhere between the two. On this hypothesis, this figure suggests that MultiRC has a large potential benefit, FEVER a small one, and E-SNLI an unclear benefit depending on how much of the predictive utility of E-SNLI rationales is due to label leakage. The results in §6 ultimately bear out these expectations.

4.2 Importance of Rationale Accuracy

We focus on MultiRC, where evaluating a non-rationale-adapted fine-tuned BERT model on human-rationalized data results in a sufficiency-accuracy of 74%, a significant improvement over the normal test accuracy of 68%. But how robust is this improvement to rationale prediction error? We examine how the sufficiency-accuracy of human rationales changes as they are corrupted by random addition, dropping, and swapping of tokens.

In this analysis, an $N\%$ drop removes $N\%$ of tokens from each rationale in the dataset, reducing recall to $100 - N$. An $N\%$ addition adds tokens numbering $N\%$ the size of each rationale, from the set of non-rationale tokens, reducing precision to $\frac{100}{100+N}$. An $N\%$ swap performs both operations, swapping $N\%$ of rationale tokens for the same number of non-rationale tokens.

The "dropped" curve in Fig. 2a shows that human rationales afford improved accuracy over the

baseline until roughly 40% of tokens have been dropped from them, suggesting that a minimum of 60% recall is needed to derive an advantage from human rationales over the full input. Per the "added" curve, adding the same number of irrelevant tokens to the rationale has a much less severe impact on accuracy, suggesting that errors of omission are significantly worse than errors of inclusion for learning from rationales.

Fig. 2b and 2c respectively show the effect of this perturbation on high- and low-sufficiency-accuracy human rationales, which constitute 74% and 26% of rationales respectively for this model. High-SA rationales follow a similar trend to the whole population, but the recall requirement is lower than Fig. 2a to exceed model accuracy with the full input (the "dropped" curve meets the blue line at 50%). In comparison, low-SA rationales demonstrate interesting properties. These rationales actually have a sabotaging effect in a quarter of cases: the model would have an accuracy of 27% with the full input, which is lowered to 0% by the presence of these rationales. Also, addition and dropping have a similar effect in mitigating this sabotage. Similar results hold on FEVER and E-SNLI except the apparent required recall is much higher (>90%) for both methods (see the appendix), indicating challenges for learning from rationales on these datasets.

In summary, our analyses inspire two general observations about learning from rationales: 1) moving away from naive accuracy (toward recall, for example) as a rationale supervision objective, and 2) focusing on useful rationales over harmful ones.

5 Methods

We propose architecture changes based on these insights. Our code is available at https://github.com/ChicagoHAI/learning-from-rationales.

5.1 Background and Baseline Models

Our training data include input tokens, their corresponding rationales, and labels. Formally, an instance is denoted as $(\boldsymbol{x}, \boldsymbol{\alpha}, y)$, where $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ is a text sequence of length L and human rationale $\boldsymbol{\alpha}$ of the same length. $\boldsymbol{\alpha}_i = 1$ indicates that token $\boldsymbol{x}_i$ is part of the rationale (and relevant for the prediction), $\boldsymbol{\alpha}_i = 0$ otherwise.

We use HuggingFace's BERT-base-uncased (Devlin et al., 2018; Wolf et al., 2020) as the basis for our experiments and analysis. Used in the standard way, BERT ignores $\boldsymbol{\alpha}$ and is fine-tuned on tuples of $(\boldsymbol{x}, y)$. This is our simplest baseline.

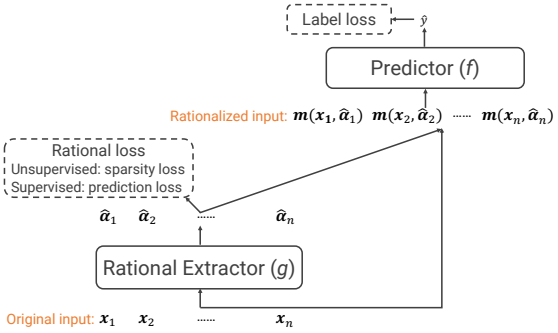


Figure 3: Illustration of our multi-task framework. Our main innovation lies in how we define rationale loss for the supervised case and the masking function m.

Rationale model. We use the rationale model of Lei et al. (2016) for both supervised and unsupervised rationale generation, in its updated BERT-to-BERT form (DeYoung et al., 2019). This model consists of two BERT modules: a rationale extractor g that generates a binary attention mask $\hat{\boldsymbol{\alpha}}$ as the rationale, and a predictor f which makes a prediction using the rationalized input via a masking function m on $\boldsymbol{x}$ and $\hat{\boldsymbol{\alpha}}$ (Fig. 3):

$$g(\boldsymbol{x}) \rightarrow \hat{\boldsymbol{\alpha}},$$
$$f(m(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})) \rightarrow \hat{y}.$$

The two components are trained in tandem. In the unsupervised scenario, the joint objective function consists of a prediction loss term and a rationale sparsity term, encouraging the model to retain only those tokens in $\boldsymbol{x}$ that are necessary for accurate prediction:

$$\mathcal{L}_u = \mathcal{L}_p(y, \hat{y}) + \lambda_{sp}||\hat{\boldsymbol{\alpha}}||,$$

where $\mathcal{L}_p$ is typically cross entropy.

In the supervised scenario, given a human rationale $\boldsymbol{\alpha}$, we replace the sparsity objective with a rationale supervision objective:

$$\mathcal{L}_{su} = \mathcal{L}_p(y, \hat{y}) + \frac{\lambda_{su}}{L}\sum_{i=1}^{L}\mathcal{L}_p(\boldsymbol{\alpha}_i, \hat{\boldsymbol{\alpha}}_i),$$

where λ_{su} is a hyperparameter that controls the weight of rationale loss compared to label loss.

Each of these scenarios represents a baseline for our experiment. We refer to the unsupervised version as *unsupervised rationale model*, and the supervised version as *supervised rationale model*.

Implementation details. The original Lei et al. (2016) model generates binary rationales by Bernoulli sampling from continuous probability values produced by the generator, and uses the REINFORCE algorithm (Williams, 1992) to prop-

agate approximate gradients through this non-differentiable operation.

We instead use Gumbel Softmax (Jang et al., 2017) to generate differentiable approximate binary rationale masks. In this framework, the generator produces logits $\boldsymbol{z_i}$ to which are added random noise $G \sim Gumbel(0, 1)$, before applying a softmax to produce class probabilities $\boldsymbol{c_i}$. This approximates a discrete distribution parameterized by $e^{\boldsymbol{z_i}}$. We then use the positive class probability $\boldsymbol{c}_i^1$ as the rationale value $\hat{\boldsymbol{\alpha}}_i$.

$$\boldsymbol{c}_i = \text{softmax}(\boldsymbol{z_i} + \boldsymbol{G} \sim Gumbel(0, 1)); \hat{\boldsymbol{\alpha}}_i = \boldsymbol{c}_i^1$$

Generating stable rationales. We find it helpful as an engineering trick to pre-train the predictor layer of this model on the full input before co-training the predictor and extractor on the joint objective. This step appears to mitigate some of the issues this model has with rationale collapse, noted for example by DeYoung et al. (2019).

Given $\hat{\boldsymbol{\alpha}}_i$, we mask non-rationale tokens by multiplicatively substituting the [MASK] token vector across their vector representations, analogously to what is done during the MASK-LM pretraining of the BERT model:

$$m_s(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}}_i) = \hat{\boldsymbol{\alpha}}_i \cdot \boldsymbol{e}_i + (1 - \hat{\boldsymbol{\alpha}}_i) \cdot \boldsymbol{e}_{\text{[MASK]}},$$

where $\boldsymbol{e}_i$ represents the embedding associated with $\boldsymbol{x}_i$ and $\boldsymbol{e}_{\text{[MASK]}}$ is the embedding for the [MASK] token. We never mask special tokens [CLS] or [SEP], and we set $\hat{\boldsymbol{\alpha}}_i = 1$ for the query in MultiRC and FEVER as well because the query is always part of human rationales in these two datasets.

5.2 Learning from Human Rationales

Inspired by the analysis in §4, we propose four strategies for improving the efficacy of learning from rationales: 1) tuning class weights for rationale supervision; 2) enforcing sentence-level rationalization; 3) using non-occluding "importance embeddings"; and 4) selectively supervising only rationales with high sufficiency-accuracy. The first three are designed to loosen the supervision's dependence on flat tokenwise accuracy, while the last tries to operationalize our observations about helpful versus non-helpful rationales.

Class weights. Rationales may become more effective enablers of model prediction accuracy at different balances of precision and recall. We can adjust this balance simply by using differing weights to positive and negative classes in rationale supervision:

$$\mathcal{L}_w = \mathcal{L}_p(y, \hat{y}) + \frac{1}{L}\sum_{i=1}^{L}(1 + \lambda_{su}^1 \alpha_i)\mathcal{L}_p(\alpha_i, \hat{\alpha}_i),$$

where λ_{su}^1 controls the relative weight of rationale vs. non-rationale tokens. In particular, as we will discuss in §4, we find that increased recall is associated with increased model accuracy. Thus, we explore several values for λ_{su}^1 in our experiment to encourage higher recall.

Sentence-level rationalization. Another divergence from strict token-wise accuracy is to rationalize at the sentence rather than the token level. Given a function $sent$ mapping a token x_i to its corresponding sentence s consisting of tokens $\{..., x_i, ...\}$, we average token-level logits $\boldsymbol{z}_i$ across each sentence to produce a binary mask at the sentence level and then propagate that mask value to all sentence tokens:

$$\hat{\boldsymbol{\alpha}}_i = \hat{\boldsymbol{\alpha}}^s_{sent(i)},$$

where $\boldsymbol{z}^s = \frac{1}{|\{i|sent(i)=s\}|}\sum_{\{i|sent(i)=s\}} \boldsymbol{z}_i$ is used to generate $\hat{\boldsymbol{\alpha}}^s_{sent(i)}$.

Importance embeddings. Another way to mitigate the impact of false negatives in predicted rationales is for these negatives to still remain visible to the predictor. This variant uses additive embeddings for rationalization rather than occluding masks, using a two-element embedding layer $\boldsymbol{e}$ constituting one embedding for rationale tokens and one for nonrationale tokens, added to the input vectors according to the predicted rationale. This way, input tokens are tagged as important or unimportant, but the predictor f has the freedom to learn how to engage with these tags for maximum label accuracy, rather than being fully blinded to "unimportant" tokens.

$$m_e(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}}_i) = \boldsymbol{e}_i + (1 - \hat{\boldsymbol{\alpha}}_i) \cdot \boldsymbol{e}_{\text{non-rationale}} + \hat{\boldsymbol{\alpha}}_i \cdot \boldsymbol{e}_{\text{rationale}}.$$

An important drawback of this approach is that the predictor now has access to the full input instead of only the rationalized input, so these rationales provide a weak guarantee that important tokens are actually used to make predictions. This method also represents a large distribution shift from full text, so we find it necessary to calibrate the predictor using human rationales, as described in Fig. 1b.

Selective supervision. Our fourth modification attempts to improve rationale prediction performance on high-sufficiency-accuracy rationales by selectively supervising only on human rationales with this property, ignoring those where human ratio-

Dataset	Model	Acc.	Rationale prediction			Human Suff. Acc.	Methods			
			F1	Prec.	Rec.		Masking	Granularity	Pos. class weight	Selective supervision
MultiRC	**BERT baseline**	68.1	-	-	-	73.9	-	Tokens	-	-
	Unsupervised rationale model	67.2	22.2	18.5	27.9	71.2	[MASK]	Tokens	-	-
	Supervised rationale model	67.0	46.5	41.5	52.9	70.8	[MASK]	Tokens	1.0	No
	Best overall model	**71.2**	**57.1**	**44.9**	**78.4**	**74.5**	Embeddings	Sentences	5.0	No
FEVER	**BERT baseline**	90.2	-	-	-	89.4	-	Tokens	-	-
	Unsupervised rationale model	88.3	22.6	20.5	25.1	88.7	[MASK]	Tokens	-	-
	Supervised rationale model	90.7	68.4	61.7	76.7	91.1	[MASK]	Tokens	1.0	No
	Best overall model	**91.5**	**81.2**	**83.5**	**79.1**	**91.6**	Embeddings	Sentences	1.0	No
E-SNLI	**BERT baseline**	89.7	-	-	-	73.9	-	Tokens	-	-
	Unsupervised rationale model	88.9	40.6	28.2	72.6	85.0	[MASK]	Tokens	-	-
	Supervised rationale model	87.8	58.7	**47.7**	76.0	89.4	[MASK]	Tokens	1.0	No
	Best overall model	**90.1**	**59.6**	45.5	**86.2**	**92.3**	Embeddings	Tokens	3.0	No

Table 3: Best-performing model variant compared to baseline models.

nales do not allow a correct prediction.

Specifically, for every training batch, we use the true human rationales $\boldsymbol{\alpha}$ as an input mask for the BERT predictor to get the HSA for each document. HSA then serves as a weight on the human rationale supervision during the main training batch:

$$\mathcal{L}_{ss}=\mathcal{L}_p(y,\hat{y})+I(y=f(m(\boldsymbol{x},\boldsymbol{\alpha})))\frac{\lambda_{su}}{L}\sum_{i=1}^{L}\mathcal{L}_p(\boldsymbol{\alpha}_i,\hat{\boldsymbol{\alpha}}_i).$$

By weighting supervision this way, we hope to ignore low-quality human rationales during training and focus instead on those that enable good accuracy.

6 Results

6.1 Experiment Setup

Our goal in this experiment is to understand the impact of our four proposed model/training modifications. We do this with a comprehensive scan: We try three positive rationale supervision class weights λ_{su}^1 ($\{0, 2, 4\}$), and toggle sentence-level rationalization, importance embedding, selective supervision on and off. In addition, we vary rationale supervision loss weight λ_{su} in $\{0.5, 1, 2\}$. This resulted in 72 models for MultiRC and FEVER, and 36 models for E-SNLI (for which sentence-level rationalization is not applicable).

The best resultant model is our *best overall model*. The best model with $\lambda_{su^1} = 1$ (i.e., identical class weights for human rationales) and no other learning strategy enabled is our baseline *supervised rationale model*. We additionally train three *unsupervised rationale models* with sparsity weights 0.15, 0.25, and 0.35, selecting as representative the one which produced the sparsest rationales while maintaining a reasonable level of accuracy (because in this architecture, there is invariably a trade-off between accuracy and sparsity).

To evaluate the performance of our models, we consider both accuracy of the predicted labels ($\hat{y}$) and performance of rationale prediction in terms of F1, precision, and recall. We use Pytorch Lightning (Falcon et al., 2019) for training with a learning rate of 2e-5 and gradient accumulation over 10 batches for all models. Early stopping was based on validation set loss with a patience of 3, evaluated every fifth of an epoch. Training was performed on two 24G NVidia TITAN RTX GPUs.

6.2 Model Performance

Table 3 compares our best overall model against the baselines, and presents the learning strategies used in the models.

Prediction accuracy. For MultiRC, this best model includes every proposed modification (sentence-level rationalization, importance embeddings, class weights) except for selective supervision, and yields a 3-point improvement from the baseline accuracy of 68.1% to 71.2%. We observe a more modest improvement on FEVER, with the best model using sentence-level rationalization and importance embeddings, and scoring a 1-point improvement from 90.2% to 91.5%. We note, however, that this approaches the accuracy of the model with access to a human rationale oracle (91.6%). Finally, we observe a tiny improvement of 0.4% on E-SNLI, though our proposed methods do improve upon the baselines of unsupervised and supervised rationale model, which causes a performance drop.

A McNemar's significance test with Bonferroni correction between the best and baseline model finds that the accuracy improvement is significant for MultiRC and FEVER (p =2e-7 and 3e-6 respectively) and not significant for E-SNLI ($p = 0.1$). The limited improvement in E-SNLI echos the performance drop in Fig. 1a without adaptation, suggesting that human rationales in this dataset are too idiosyncratic to improve model performance.

Factor analysis. We use regression analysis to

Method	Coefficients		
	MultiRC	FEVER	E-SNLI
Sentences	.015***	.001	-
Class weights	.017***	.007***	.005
Importance embeddings	.012***	.006***	-.010**
Selective supervision	0.004	-.006***	-.032***

Table 4: Regression coefficients for effect each proposed method on overall prediction accuracy

Dataset	Sel. Sup.	Acc.	F1.	
			High-HSA	Low-HSA
MultiRC	**No**	**71.2**	**59.3**	**57.2**
	Yes	71.0	56.2	54.1
FEVER	**No**	**91.5**	**79.0**	**72.5**
	Yes	90.6	61.2	57.0
E-SNLI	**No**	**90.1**	**61.2**	**48.0**
	Yes	88.8	49.0	44.9

Table 5: Label accuracy and predicted rationale F1 for high- versus low-HSA examples.

understand the impact of the different modifications on model accuracy. Table 4 suggests that rationale class weighting has the highest positive effect on accuracy across datasets. Importance embeddings have a positive effect for MultiRC and FEVER and a negative effect for E-SNLI, while sentence-level rationalization improves only MultiRC.

Selective supervision is found to have a non-existant or negative effect across all three datasets. Table 5 details this result, showing model accuracy and rationale performance for the best model with (yes) vs. without (no) selective supervision. If our method succeeded, F1 for high-HSA examples would increase from the "No" to the "Yes" models and remain flat or decrease for low-HSA examples. Indeed, we observe lower rationale F1 for low-HSA examples, but the rationale F1 also drops substantially for high-HSA examples, possibly because of the reduced available training data.

Rationale performance. Although our modifications are designed to improve label prediction performance, they also improve rationale prediction performance in most cases. The only exception is the reduced precision in E-SNLI compared to the supervised rationale model.

6.3 Qualitative Analysis

Table 6 shows three examples, each drawn from a different dataset, to illustrate different outcomes. For each example, we show the human rationale and predicted rationales for both the baseline supervised rationale model and our best overall model. Incorrect predictions are colored red.

Example 6a shows an instance sampled from MultiRC where our best model, with higher recall and sentence-level rationalization, more successfully captures the (sufficient) information present in the human rationale, allowing for a correct prediction where the supervised rationale model fails.

Example 6b presents a contrasting example from the FEVER dataset. The human rationale omits important context, that Legendary Entertainment is a subsidiary of Wanda Group, making it harder to infer that it is *not* a subsidiary of Warner Bros. Our best model succeeds at capturing this snippet in its rationale, but still predicts the incorrect label, illustrating that a sufficient (for humans) rationale does not always produce a correct label.

Finally, example 6c shows a case where the baseline supervised rationale model succeeds while our best model fails. This is a hard-to-interpret example, mainly a demonstration of the limitations of rationales as an explanatory device for certain kinds of task. This begs a question: how relevant are rationales as an explanation or learning mechanism when models like GPT-3 (Brown et al., 2020) are increasingly capable of human-level *natural language* explanations (Table 7)?

Our position is that however an explanation is presented, meaning is still localized within text, so rationales can still serve as a useful interface for scrutinizing or controlling model logic, even if they require additional translation to be comprehensible to humans. Works that hybridize the two ideas such as Zhao and Vydiswaran (2020) may represent a good way of resolving this issue.

7 Discussion

The analysis in section §4 explores the limits of potential improvement from learning from rationales. It suggests two insights toward improved learning from rationales: 1) that insofar as they boost model accuracy, not all human rationale tokens are equally valuable, e.g., with false positives causing less degradation than false negatives; and 2) we could in principle boost label accuracy with good rationale accuracy on useful (high-SA) rationales and low accuracy on useless (low-SA) ones.

We exploit these two insights with four modifications to the baseline architecture. Three of these diverge from flat rationale supervision accuracy: rationale supervision class weighting, sentence-level rationalization, and importance embeddings. The last, selective supervision, pursues utility-discriminative weighting during model training.

Human rationale	Baseline supervised rationale	Best model
(A) MultiRC: Best model beats supervised baseline		
[CLS] there have been many organisms that have lived in earths past . only a tiny number of them became fossils . still , scientists learn a lot from fossils . fossils are our best clues about the history of life on earth . fossils provide evidence about life on earth . they tell us that life on earth has changed over time . fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . [SEP] what can we tell about former living organisms from fossils ? | | how they adapted [SEP]	[CLS] there have been many organisms that have lived in earths past . only a tiny number of them became fossils . still , scientists learn a lot from fossils . fossils are our best clues about the history of life on earth . fossils provide evidence about life on earth . they tell us that life on earth has changed over time . fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . [SEP] what can we tell about former living organisms from fossils ? | | how they adapted [SEP]	[CLS] there have been many organisms that have lived in earths past . only a tiny number of them became fossils . still , scientists learn a lot from fossils . fossils are our best clues about the history of life on earth . fossils provide evidence about life on earth . they tell us that life on earth has changed over time . fossils in younger rocks look like animals and plants that are living today . fossils in older rocks are less like living organisms . fossils can tell us about where the organism lived . was it land or marine ? fossils can even tell us if the water was shallow or deep . fossils can even provide clues to ancient climates . [SEP] what can we tell about former living organisms from fossils ? | | how they adapted [SEP]
Prediction: False	**Prediction: True**	**Prediction: False**
(B) FEVER: Human rationale is insufficient		
[CLS] legendary entertainment - lrb - also known as legendary pictures or legendary - rrb - is an american media company based in burbank , california . the company was founded by thomas tull in 2000 and in 2005 , concluded an agreement to co - produce and co - finance films with warner bros . , and began a similar arrangement with universal studios in 2014 . since 2016 , legendary has been a subsidiary of the chinese conglomerate wanda group . [SEP] legendary entertainment is a subsidiary of warner bros pictures . [SEP]	[CLS] legendary entertainment - lrb - also known as legendary pictures or legendary - rrb - is an american media company based in burbank , california . the company was founded by thomas tull in 2000 and in 2005 , concluded an agreement to co - produce and co - finance films with warner bros . , and began a similar arrangement with universal studios in 2014 . since 2016 , legendary has been a subsidiary of the chinese conglomerate wanda group . [SEP] legendary entertainment is a subsidiary of warner bros pictures . [SEP]	[CLS] legendary entertainment - lrb - also known as legendary pictures or legendary - rrb - is an american media company based in burbank , california . the company was founded by thomas tull in 2000 and in 2005 , concluded an agreement to co - produce and co - finance films with warner bros . , and began a similar arrangement with universal studios in 2014 . since 2016 , legendary has been a subsidiary of the chinese conglomerate wanda group . [SEP] legendary entertainment is a subsidiary of warner bros pictures . [SEP]
Prediction: Supports	**Prediction: Supports**	**Prediction: Supports**
(C) E-SNLI: Supervised baseline beats best model		
[CLS] a big dog catches a ball on his nose [SEP] a big dog is sitting down while trying to catch a ball [SEP]	[CLS] a big dog catches a ball on his nose [SEP] a big dog is sitting down while trying to catch a ball [SEP]	[CLS] a big dog catches a ball on his nose [SEP] a big dog is sitting down while trying to catch a ball [SEP]
Prediction: Neutral	**Prediction: Neutral**	**Prediction: Contradiction**

Table 6: Examples of human, supervised baseline, and best model rationales and predictions.

Source	Natural language explanation
Human	There is no indication that the dog is sitting down while playing catch on his nose.
Human	A dog can catch a ball by not to sitting down.
GPT-3	The entailment of this sentence is that the dog is sitting down, and the contradiction would be if the dog was standing up. This sentence is neutral, meaning it doesn't entail or contradict anything.

Table 7: Examples of natural language explanations for the "neutral" prediction on E-SNLI example from Table 6c. See Appendix §D for GPT-3 prompt details.

Taken together, our proposed methods yield a substantial 3% improvement over baseline performance for MultiRC, a 1% improvement on FEVER, and a tiny .4% improvement on E-SNLI, mirroring the potential improvements observed in the analysis. We find that all three token supervision methods are useful in achieving this, while selective supervision has a marginal or negative effect.

In summary, our results support the potential for learning from rationales in certain datasets, and demonstrate the importance of understanding the properties of human rationales to properly exploit them for this purpose. We believe that these two insights are useful steps towards effective learning from rationales, and could yield even greater improvements if operationalized optimally.

Limitation. A limitation of our analysis is that all three datasets are document-query style reading comprehension tasks, as opposed to, e.g., sentiment analysis. Because of the popularity of this type of task in NLP benchmarks, this type of dataset represents a majority of what is available in the ERASER collection (DeYoung et al., 2019). By contrast, sentiment is often scattered throughout a text, so human rationales for sentiment are likely to contain redundant signal, which could impact their predictive utility. We leave a more comprehensive survey of NLP tasks for future work.

Acknowledgments. We thank anonymous reviewers for their feedback, and members of the Chicago Human+AI Lab for their insightful suggestions. This work is supported in part by research awards from Amazon, IBM, Salesforce, and NSF IIS-2126602.

References

Diego Antognini and Boi Faltings. 2021. Rationalization through Concepts. *arXiv:2105.04837 [cs]*. ArXiv: 2105.04837.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2020. Interpretable Neural Predictions with Differentiable Binary Variables. *arXiv:1905.08160 [cs]*. ArXiv: 1905.08160.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Proceedings of NeurIPS*.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

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

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant Rationalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1448–1458. PMLR. ISSN: 2640-3498.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv:1911.03429 [cs]*. ArXiv: 1911.03429.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active Learning by Labeling Features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 81–90, Singapore. Association for Computational Linguistics.

William Falcon et al. 2019. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of ACL*.

Peter Hase and Mohit Bansal. 2021. When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data. *arXiv:2102.02201 [cs]*. 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:2010.04119 [cs]*. ArXiv: 2010.04119.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *arXiv:2006.01067 [cs]*. ArXiv: 2006.01067.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. *arXiv:1611.01144 [cs, stat]*. ArXiv: 1611.01144.

Sahil Jayaram and Emily Allaway. 2021. Human Rationales as Attribution Priors for Explainable Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. 2021. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. *arXiv:2103.01890 [cs, stat]*. ArXiv: 2103.01890.

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

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 126–137, Atlanta Georgia USA. ACM.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *arXiv:2104.15135 [cs]*. ArXiv: 2104.15135.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. *arXiv:2005.00652 [cs]*. ArXiv: 2005.00652.

Gregory Plumb, Maruan Al-Shedivat, Angel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing Black-box Models for Improved Interpretability. *arXiv:1902.06787 [cs, stat]*. ArXiv: 1902.06787.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR. ISSN: 2640-3498.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *arXiv preprint arXiv:1703.03717*. 00000.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to Spurious Correlations via Human Annotations. *arXiv:2007.06661 [cs, stat]*. ArXiv: 2007.06661.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL*.

Giulia Vilone and Luca Longo. 2020. Explainable Artificial Intelligence: a Systematic Review. *arXiv:2006.00093 [cs]*. ArXiv: 2006.00093.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing Human-AI Collaboration for Generating Free-Text Explanations. *arXiv:2112.08674 [cs]*. ArXiv: 2112.08674.

Sarah Wiegreffe and Ana Marasović. 2021. Teach Me to Explain: A Review of Datasets for Explainable NLP. *arXiv:2102.12060 [cs]*. ArXiv: 2102.12060.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining Neural Networks with Compositional Explanations. *arXiv:2103.10415 [cs]*. ArXiv: 2103.10415.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S. Jaakkola. 2019. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. *arXiv preprint*. ArXiv: 1910.13294.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi S. Jaakkola. 2021. Understanding Interlocking Dynamics of Cooperative Rationalization. *arXiv:2110.13880 [cs]*. ArXiv: 2110.13880.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2020. LIREx: Augmenting Language Inference with Relevant Explanation. *arXiv:2012.09157 [cs]*. ArXiv: 2012.09157.

Dataset	Method	Role	Accuracy	Rationale prediction			Human Suff. Acc.
				F1	Precision	Recall	
MultiRC	**Sentences**	Best with	71.2	57.1	44.9	78.4	74.5
	Sentences	Best without	70.6	41.6	27.7	84.1	75.8
	Class-weights	Best with	71.2	57.1	44.9	78.4	74.5
	Class-weights	Best without	70.8	55.2	66.1	47.4	76.5
	Importance embeddings	Best with	71.2	57.1	44.9	78.4	74.5
	Importance embeddings	Best without	71.0	53.6	39.7	82.5	75.8
	Selective supervision	Best with	71.0	53.6	39.7	82.5	75.8
	Selective supervision	Best without	71.2	57.1	44.9	78.4	74.5
FEVER	**Sentences**	Best with	91.5	81.2	83.5	79.1	91.6
	Sentences	Best without	91.3	72.4	61.3	88.5	91.6
	Class-weights	Best with	91.5	79.6	73.1	87.3	91.8
	Class-weights	Best without	91.5	81.2	83.5	79.1	91.6
	Importance embeddings	Best with	91.5	81.2	83.5	79.1	91.6
	Importance embeddings	Best without	91.4	80.0	74.9	85.9	91.8
	Selective supervision	Best with	90.6	56.4	41.4	88.6	90.4
	Selective supervision	Best without	91.5	81.2	83.5	79.1	91.6
E-SNLI	**Class-weights**	Best with	90.1	59.6	45.5	86.2	92.3
	Class-weights	Best without	89.9	62.2	55.7	70.4	92.0
	Importance embeddings	Best with	90.1	59.6	45.5	86.2	92.3
	Importance embeddings	Best without	89.9	33.5	20.2	100.0	72.5
	Selective supervision	Best with	88.8	49.0	33.2	93.4	84.0
	Selective supervision	Best without	90.1	59.6	45.5	86.2	92.3

Table 8: Comparison of best model with each proposed factor against best model without that factor.

A Detailed Factor Analysis

Table 8 compares, for each proposed method, the performance of the best model using that method and the best model not using it. The story shown here is similar to the regression analysis in Table 4, but one new insight is that the improvement in model prediction performance appears to be driven by the sentence-level rationalization method, as it cuts down on stray tokens dropped from or added to the predicted rationales.

B Rationale Perturbation on FEVER and E-SNLI

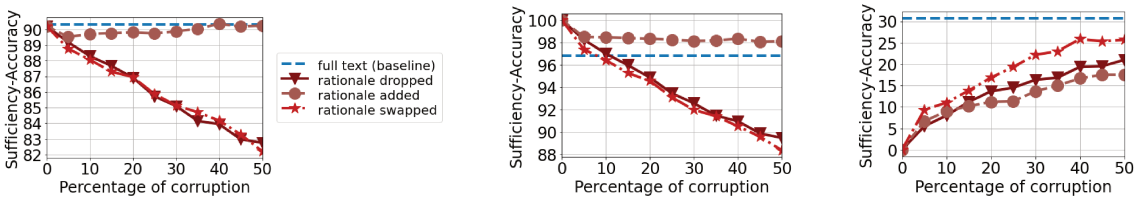
Furthering the analysis in §4.2, we extend the human rationale perturbation experiment to FEVER and E-SNLI.

Fig. 4 show the result for FEVER. Fig. 4a shows that the baseline accuracy is so high for this dataset that to match just the baseline accuracy for FEVER, we require near perfect prediction of human rationales.

Moreover, even for documents with HSA $= 1$, the model performance drops below baseline on dropping just $\sim$ 10% tokens (synonymous with rationale recall = $\sim$0.9) in Fig. 4b. Interestingly, the model performance remains consistently above the baseline when adding non-rationale tokens (synonymous with decreasing rationale precision). In comparison, the model performance for MultiRC in Fig. 2b drops below baseline after dropping $\sim$50% of the tokens.

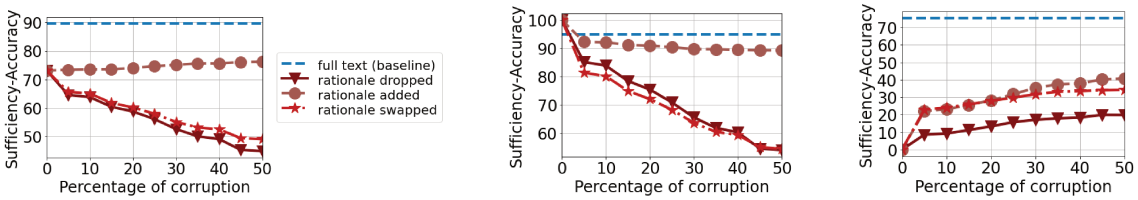
For FEVER examples with HSA $= 0$ (Fig. 4c), the model performance remains below the baseline accuracy consistently, supporting the second hypothesis in §4.2. The near-perfect need to predict rationales in FEVER may explain behind the difference in improvements of model performance between MultiRC and FEVER.

Fig. 5 covers E-SNLI. We see that the model performance decreases after dropping rationale tokens (signifying decreasing recall) and it consistently remains below the baseline. In contrast, the model performance shows a slight improvement after adding non-rationale tokens (signifying decrease in rationale precision). Moreover, for documents with HSA $= 1$, the model performance drops below baseline at $\sim$3% for dropping and swapping rationale tokens, where as the model performance plateaus with addition of non-rationale tokens. These insights highlights the substantial challenges in learning from explanations for E-SNLI.



(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 4: Performance of corrupted rationale for FEVER. Model performance drops below baseline accuracy immediately on both dropping human rationales (i.e., recall ↓) and adding non-rationale tokens (i.e., precision ↓). For HSA = 1, model performance remains consistently above baseline on adding non-rationale tokens (i.e. precision ↓)



(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 5: Performance of corrupted rationales for E-SNLI. Model performance for human rationale remains below baseline accuracy and slightly increases with addition of non-rationale tokens (i.e. precision ↓). Even for HSA = 1, model performance drops below baseline accuracy at just ∼4% corruption.

C Rationale Perturbation for Adapted Models

We perform the same perturbation analysis on calibrated model trained on both full and rationalized input, for which distribution shift from masking are less of a concern.

In Fig. 6, for MultiRC, we find that model performance plateaus with addition of non-rationale tokens and drops quickly with rationale tokens even for a calibrated model. This observation is consistent for FEVER (Fig. 7).

For E-SNLI, we find different properties using a calibrated BERT model compared to the standard BERT model show in Fig. 5a.

In contrast to MultiRC and FEVER, we find that the model performance drops more rapidly with the addition of non-rationale tokens compared to removal of rationale tokens. This is consistent for documents with HSA = 1, suggesting that for E-SNLI, rationale precision maybe more important when using a calibrated model. Similar to FEVER, we see the model performance drop below the baseline with very little corruption of rationales, echoing the need to perfectly mimic human rationalization for effective learning from rationales for this dataset.

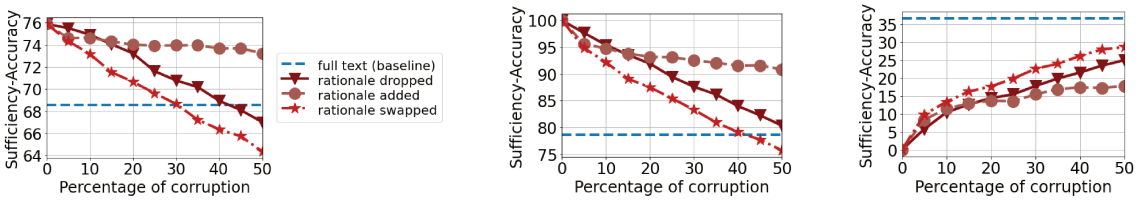
D GPT-3 Prompt

We generate a zero-shot GPT-3 (Brown et al., 2020) explanation using the Davinci model variant on the OpenAI playground[1], and a modified version of the prompt proposed by Wiegreffe et al. (2021):

> Let's explain classification decisions.
>
> A big dog catches a ball on his nose.
>
> question: A big dog is sitting down while trying to catch a ball.
>
> entailment, contradiction, or neutral?

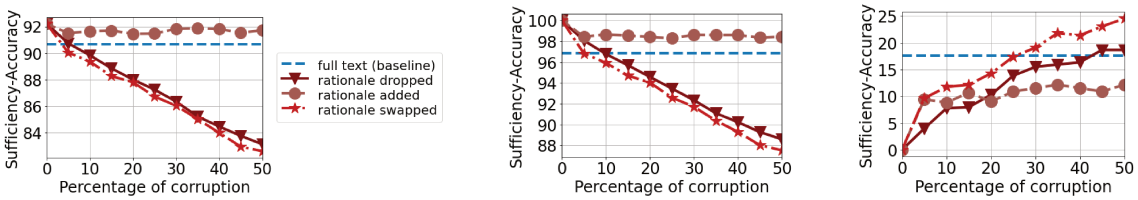
A second step prompting for an explanation is not needed, as GPT-3 gives its prediction in the form of a natural language explanation.

[1] https://beta.openai.com/playground



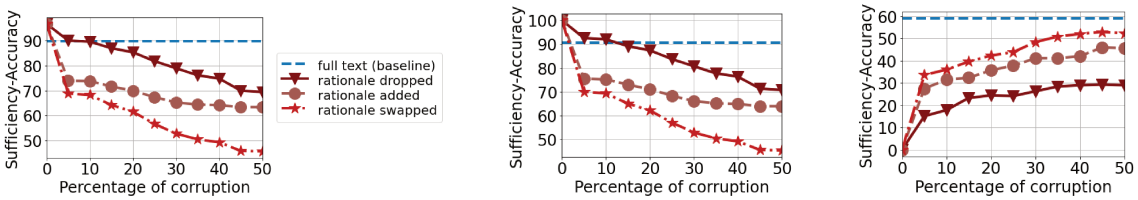
(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 6: Performance of corrupted rationales for MultiRC using a calibrated model. Model performance decreases consistently when we drop human rationales (i.e., recall ↓), where as the model performance stays high as we add non-rationale tokens (i.e., precision ↓). The impact of recall is moderated when HSA$= 1$.



(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 7: Performance of corrupted rationales for FEVER using a calibrated model. Model performance decreases quickly when we drop human rationales (i.e., recall ↓), where as the model performance remains above baseline as we add non-rationale tokens (i.e., precision ↓).



(a) All samples (b) Human sufficiency-accuracy = 1 (c) Human sufficiency-accuracy = 0

Figure 8: Performance of corrupted rationales for E-SNLI using a calibrated model. Model performance decreases quickly when we add non- rationale tokens (i.e., precision ↓), where as the model performance drops less rapidly as we drop rationale tokens (i.e., recall ↓).