Evaluating and Characterizing Human Rationales

Samuel Carton[*], **Anirudh Rathore**[*], **Chenhao Tan**
University of Colorado Boulder
{samuel.carton,anirudh.rathore,chenhao.tan}@colorado.edu

Abstract

Two main approaches for evaluating the quality of machine-generated rationales are: 1) using human rationales as a gold standard; and 2) automated metrics based on how rationales affect model behavior. An open question, however, is how human rationales fare with these automatic metrics. Analyzing a variety of datasets and models, we find that human rationales do not necessarily perform well on these metrics. To unpack this finding, we propose improved metrics to account for model-dependent baseline performance. We then propose two methods to further characterize rationale quality, one based on model retraining and one on using "fidelity curves" to reveal properties such as irrelevance and redundancy. Our work leads to actionable suggestions for evaluating and characterizing rationales.

1 Introduction

Explanations in NLP often take the form of *rationales*, subsets of input tokens that are considered important to the model's decision (DeYoung et al., 2020). As interest in explainable AI has increased, so has interest in evaluating the quality of explanatory rationales. However, this is a challenging task because it can be difficult to pin down exactly what constitutes "good" rationales for model predictions (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019).

Two main strategies that have been proposed in recent work are: 1) to view human-generated rationales as a gold standard and evaluate model-generated rationales in comparison to them; and 2) to assess the "fidelity" of a rationale to a prediction using automatic metrics.

The human-gold-standard approach views rationales as an additional form of label that can be collected alongside document-level labels. Because NLP tasks tend to involve human-generated labels, it makes intuitive sense that human-generated rationales might be considered authoritative.

When human rationales are not available, evaluations of machine rationales turn to automatic metrics. These metrics divorce rationale evaluation from an external standard, seeking instead to judge whether rationales are coherent relative to model behavior. Popular recent metrics are *sufficiency* and *comprehensiveness* (i.e., necessity), which assess whether a rationale is sufficient/necessary for a model prediction by comparing the model's behavior on the full input to its behavior on input masked according to the rationale or its complement. We use the term *fidelity* to refer jointly to sufficiency and comprehensiveness.

To the best of our knowledge, no existing work has systematically examined human rationales using these automatic metrics. However, this is an important step towards evaluating rationales because it helps characterize the disparities between the two types of approach. Are human rationales sufficient to allow models to predict human labels? Are they comprehensive? And what other insights can we gain about human rationales and fidelity metrics by performing this assessment?

In practice, both human rationales and automatic metrics can fail to work as intended (Table 1). For instance, human rationales may be insufficient because they fail to include needed information (e.g., the album title in Table 1.1), or non-comprehensive because they miss redundant-yet-relevant information (e.g., the second personal attack in Table 1.2).

By contrast, a truly sufficient rationale can be deemed insufficient due to a model not learning expected classification rules (e.g., "sits" $\sim$ "laying" in Table 1.3). While this type of failure is inevitable in machine learning, more avoidable are cases where model bias causes rationales to be evaluated incorrectly or inconsistently. For instance,

[*]Equal contribution.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 9294–9307,
November 16–20, 2020.

	Human rationale	Sufficiency	Comprehensiveness	Failure type	Dataset
1.	No Way Out is the debut studio album by ... Puff Daddy . **It was released on July 1 , 1997 , by his Bad Boy record label** . ... [SEP] **1997 was the year No Way Out was released.**	0.005	0.224	Human	FEVER
2.	**== what the f*** is your problem , b**** !!!!!!!!!!!** == why the f*** did you delete the dreamtime festival page , s******	1.0	0.001	Human	WikiAttack
3.	A man **sits** on a couch beside a colorful cushion with a pencil in his hand. [SEP] The man is **laying** down on the couch.	0.002	0.999	Metric	E-SNLI
4.	:: makes sense . have a good one .	0.971	0.0	Metric	WikiAttack

Table 1: Example rationales drawn from various datasets. Underlined tokens are rationales provided by humans. Human annotators can fail to produce faithful rationales (row 1 and 2), and fidelity metrics themselves can be misleading (row 3 and 4).

in Table 1.4, the model has learned a heavy bias toward the no-attack class (i.e., the model predicts no-attack for the empty input), so an empty rationale for a no-attack prediction is deemed perfectly sufficient yet entirely noncomprehensive.

To investigate the empirical properties of human rationales and automatic metrics, we analyze the fidelity of human rationales across six datasets. We show that human rationales do not necessarily have high sufficiency or comprehensiveness based on automatic metrics, and their fidelity varies greatly from model to model and class to class.

We propose extensions to existing fidelity metrics and develop novel methods to further characterize the quality of human rationales. First, we note that fidelity is highly model-dependent, and that model behavior can result in misleading fidelity results. We propose a normalization procedure to allow for fair comparison of these metrics across models, classes, and datasets. We show that this normalization helps contextualize fidelity results by accounting for baseline model behavior.

Second, we evaluate model accuracy on full vs. rationale-only data, linking typical output-sufficiency to performance outcomes (i.e., *accuracy-sufficiency*). We examine the effect of allowing models to adapt to rationale-only data during training, drawing a distinction between a rationale's "incidental" fidelity and its "potential" fidelity to a model. We analyze the effect of these two interventions and discuss their implications for evaluation of (and learning from) human rationales.

Finally, we introduce the idea of "fidelity curves", which examine how sufficiency and comprehensiveness degrade as tokens are randomly occluded from a rationale. We discuss how the shapes of these curves can be used to infer fine-grained attributes about rationales, such as the extent to which they contain redundant or highly interdependent tokens. We find that rationales in our datasets vary greatly in their level of irrelevancy, redundancy, and mutual dependence. We find that our three classification tasks exhibit less dependence and more redundancy in their rationales than our three document/query-style tasks.

Evaluating rationales is a significant challenge. We argue that in order to be confident in either human rationales or automatic fidelity metrics, we have to understand how these two evaluation approaches interact with one another, and what caveats they can reveal about each other. Our analyses lead to the following actionable implications:

- Fidelity metrics are highly model-dependent and should be normalized to assist interpretation.
- Models trained on rationale-only data can provide accuracy-based metrics to complement the "incidental" metrics.
- "Fidelity curves" provide a novel way to infer fine-grained qualities about rationales, such as irrelevance and redundancy.

2 Datasets

The goal of this paper is to evaluate and characterize human rationales. We analyze six datasets, four drawn from the ERASER collection (DeYoung et al., 2020), and two from other sources. They consist of three single-text classification tasks and three document/query-style tasks where it is important to understand the relations between texts.

For each dataset, the human rationales have a qualitative *expected comprehensiveness* based on whether, by construction or design, they are intended to contain all pertinent information for their respective prediction task. Four of our six datasets are expected to have comprehensive rationales.

- **WikiAttack** (Carton et al., 2018). A classification dataset of 115,859 Wikipedia revision com-

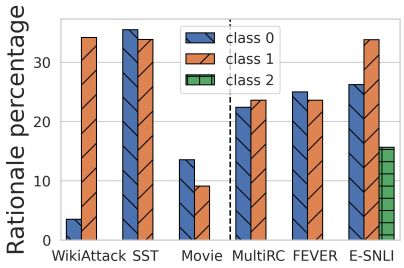
Dataset	Text length	Task type	Rationale				
			Length	Ratio	Comprehensive	Granularity	Class asymmetry
WikiAttack	51.8	classification	6.5	19.1%	✓	Token	✓
SST	19.3	classification	6.5	34.6%	✓	Token	✗
Movie	774.3	classification	82.4	11.3%	✗	Token	✗
MultiRC	321.7	document/query-style	69.8	22.9%	✓	Sentence	✗
FEVER	320.7	document/query-style	53.6	24.0%	✗	Sentence	✗
E-SNLI	21.3	document/query-style	5.0	25.2%	✓	Token	✓

Table 2: Basic statistics. Dataset rationales exhibit a range of average rationale-to-text ratios, expected comprehensivenesses, granularities, and class asymmetries.

ments labeled for presence of personal attacks by Wulczyn et al. (2017) and augmented with 1,049 human rationales by Carton et al. (2018). The rationales in this dataset are expected to be comprehensive, as labelers were asked to identify all personal attacks in each text.

- **Stanford Sentiment Treebank (SST)** (Socher et al., 2013). A classification dataset of 9,620 movie review snippets annotated for positive/negative sentiment at every syntactic tree node. We flatten these into rationales using a heuristic algorithm (see the appendix). The rationales are expected to be comprehensive, as they contain all high-sentiment phrases.
- **Movie** (Zaidan and Eisner, 2008). A classification dataset of 2,000 movie reviews labeled with rationales. The rationales are not necessarily comprehensive, as annotators were not instructed to identify all evidence.
- **MultiRC** (Khashabi et al., 2018) A reading comprehension dataset of 32,091 document-question-answer triplets that are true or false. Rationales are expected to be comprehensive as they each 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 sub-snippet (and the claim itself), and are not expected to be comprehensive as they may not cover all pertinent information.
- **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. Explanations for this dataset are expected to be comprehensive as the texts are short and labelers were instructed to identify all relevant tokens.

Table 2 shows the basic statistics of each dataset. Significant variation exists between datasets in rationale length and rationale percentage. For example, rationales only cover 11.3% of the words in Movie, consistent with our expectation of non-comprehensiveness. We also report rationale granularity, whether annotations were provided at the token or sentence level, and class asymmetry, whether rationale lengths vary significantly between classes. For the purpose of this analysis, tokenization is provided by the individual dataset sources, so we simply split texts by whitespace.



WikiAttack	0: no-attack, 1: personal-attack
SST	0: negative, 1: positive
Movie	0: negative, 1: positive
MultiRC	0: false, 1: true
FEVER	0: refutes, 1: supports
E-SNLI	0: contradiction, 1: entailment, 2: neutral

Figure 1: Percentage of rationales by class. Significant variations exist in WikiAttack and E-SNLI.

Fig. 1 shows class asymmetry in rationale percentages. For WikiAttack, labelers were asked to highlight personal attacks, and thus evidence for the no-attack class comes in the form of no highlighted tokens. This results in a situation where rationales for no-attack examples constitute less than 5% on average, while they constitute 35% of personal-attack examples. Significant variation between classes also exists in E-SNLI: entailment contains close to 40% of tokens as rationales, but neutral merely consists of 16% — another case of evidence through absence (negative evidence).

3 Evaluating Human Rationales

Popular automatic metrics for evaluating machine-generated rationales are *sufficiency* and *comprehensiveness*, articulated by Yu et al. (2019) and em-

ployed in the ERASER benchmark (DeYoung et al., 2020). *Sufficiency* measures how well rationales can provide the same prediction as using full information, while *comprehensiveness* measures how well rationales include all relevant information.

It remains an open question whether human-generated rationales have good sufficiency and comprehensiveness. We find that this is in fact not necessarily the case. This result reveals a contradiction in the evaluation of machine-generated rationales: human-generated rationales are used as a gold standard, but being similar to human-generated rationales may not lead to high sufficiency and comprehensiveness. Another important observation from our experiments is that there exists significant variation between datasets and classes within the same dataset.

3.1 Formal Definitions & Experiment Setup

A rationale is *sufficient* if it contains enough information to allow the model to make a prediction close to what it would make with full information. Formally, we represent rationales as a binary mask $\boldsymbol{\alpha}$ over the input $\boldsymbol{x}$ that indicates whether each token belongs to the rationale or not (1 to include, 0 to exclude). The sufficiency of rationales for a given prediction $\hat{y}$ is based on the difference in class probability between using full information and using only the rationale:

$$\text{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = 1 - \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \boldsymbol{\alpha})), \quad (1)$$

where $\hat{y} = \arg\max_y p(y|\boldsymbol{x})$. Note that we use the reverse of the difference so that higher sufficiency indicates faithful rationales. We also enforce the difference in class probability to be above 0, which differs from DeYoung et al. (2020).[1] This operation bounds sufficiency to between 0 and 1.

Comprehensiveness (i.e., necessity) captures the extent to which a rationale is needed for a prediction, by assessing the model's prediction on the complement of the rationale ($\mathbf{1} - \boldsymbol{\alpha}$). For a highly comprehensive explanation, the model's prediction on its complement should differ greatly from its prediction on the full information. As above, we enforce this value to be bounded between 0 and 1:

$$\text{Comp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \mathbf{1} - \boldsymbol{\alpha})). \quad (2)$$

Our definitions entail that a faithful rationale should have both high sufficiency and comprehensiveness.

[1] Arguably, the sufficiency metric should not go above 1 no matter how good the rationales are. That said, our results demonstrate similar qualitative trends from the definitions without the max operation. See the appendix.

Implicit in the definition of sufficiency and comprehensiveness is a dependence on the properties of the underlying model. To study the relationship between model property and human rationale fidelity, we experiment with a range of models: logistic regression, random forests, LSTM (Hochreiter and Schmidhuber, 1997) and RoBERTa (Liu et al., 2019). We use the same train/dev/test splits as in the original datasets. We report the resulting model with the best validation accuracy in the main paper. To apply rationale masking, we simply remove the tokens which correspond with 0s in the rationale mask. See the supplementary material for implementation details. Our code is available at `https://github.com/BoulderDS/evaluating-human-rationales`.

3.2 Overall Results

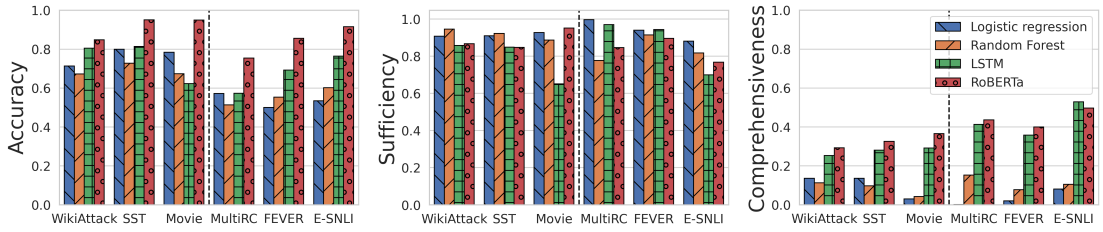
Fig. 2a shows the accuracy of our models on each dataset. As expected, RoBERTa shows the best performance followed generally by LSTM, then random forest and logistic regression. The only exception is Movie, where LSTM models struggle with the long texts (774 tokens on average) due to the limited dataset size and vanishing gradients.

We find that **human rationales do not necessarily have high sufficiency and comprehensiveness**. Moreover, human-generated rationales obtain weaker sufficiency in highly accurate models (Fig. 2b). In fact, human rationales have lower sufficiency in RoBERTa than logistic regression or random forest in five of six datasets. This finding demonstrates that the sufficiency of an explanation can be inversely correlated with model performance, which is a problem for comparing explanation methods across different models.

By contrast, strong models show better comprehensiveness scores for human rationales (Fig. 2c), with values ranging from 0.3 to 0.5 for RoBERTa. E-SNLI demonstrates the highest comprehensiveness in this model while Movie and MultiRC, both expected to be non-comprehensive, respectively achieve the 2nd and 4th highest comprehensiveness, in defiance of our expectations.

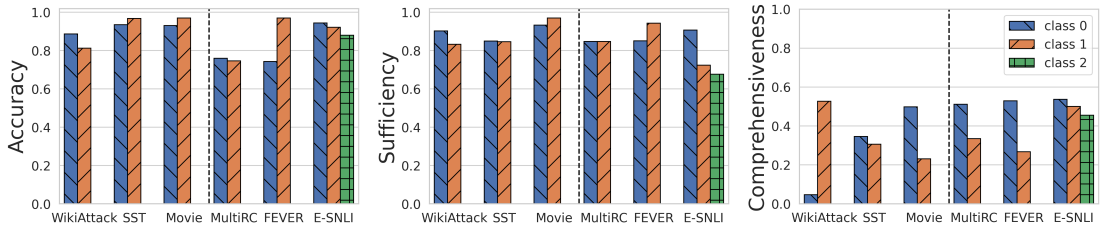
Moving forward, we focus on RoBERTa as it is the most accurate and represents the industry standard for general NLP.

Classes matter. Breaking down fidelity by class reveals further nuances. Fig. 3b shows that sufficiency is mostly even between classes, though significant differences exist for E-SNLI. Surpris-



(a) Accuracy (b) Sufficiency (c) Comprehensiveness

Figure 2: Accuracy, sufficiency, and comprehensiveness of human rationales with different models. While RoBERTa performs significantly better in all datasets in accuracy, it is rarely the best in sufficiency. In comparison, human rationales tend to have abysmal comprehensiveness with classic models.



(a) Accuracy by class (b) Sufficiency by class (c) Comprehensiveness by class

Figure 3: Accuracy, sufficiency, and comprehensiveness of human rationales grouped by class for RoBERTa. While sufficiency is relatively stable across classes, we observe dramatic differences between classes in comprehensiveness (e.g., WikiAttack and Movie).

ingly, in WikiAttack, sufficiency is higher in the no-attack class where there are a small number of tokens in human rationales.

The evenness in sufficiency is not mirrored in comprehensiveness (Fig. 3c), which differs wildly from class to class for different datasets. The most extreme case is WikiAttack, where by design the "rationale" for a no-attack comment is for nothing to be highlighted. The comprehensiveness of these empty rationales is correspondingly low. Interestingly, E-SNLI demonstrates a relatively even spread of comprehensiveness across classes despite its class-asymmetric rationale lengths.

Movie, MultiRC, and FEVER all show large class discrepancies in comprehensiveness despite having similar-length rationales across classes. In FEVER, for example, this means that removing the identified evidence for a "refutes" outcome tends to have a higher impact on the model prediction than for "support" outcomes. This could be due to task semantics (e.g., that refuting evidence is generally more unique than supporting evidence), or model bias (e.g., that the model tends to predict "supports" by default and therefore is less affected by removing the rationales for this outcome).

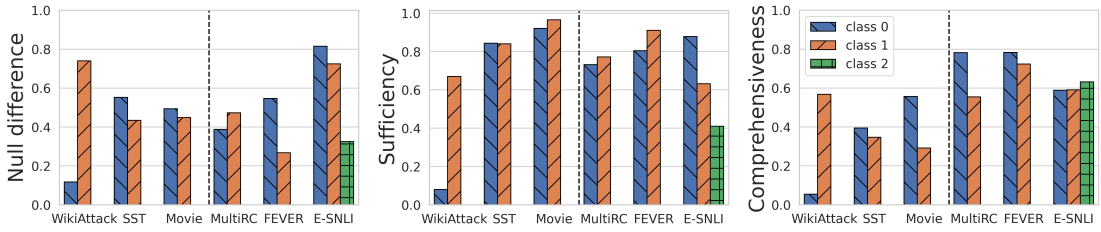
4 Normalizing Sufficiency and Comprehensiveness

Human rationales do not necessarily have high fidelity, suggesting that either human rationales or evaluation metrics may be problematic. We start by rethinking the fidelity metrics in this section and will propose novel methods to characterize human rationales in §5.

A salient observation in Fig. 2 is that sufficiency and comprehensiveness are in completely separate value ranges, although they are both theoretically bounded between 0 and 1. To properly interpret these numbers, we need to establish a baseline for them. We do so by defining a "*null difference*", the difference in output between the model operating on full information vs. no information (i.e., the empty input). This value is equivalent to (the complement of) the sufficiency of an all-zero (empty) rationale mask, or the comprehensiveness of an all-one mask.

Null difference is an intrinsic value for a given model and dataset, and depends on the class balance of the dataset, the bias term(s) learned by the model, and the calibration of output probability. It serves as a baseline value in the sense that no rationale should be much less sufficient than an all-zero rationale or much more comprehensive than an all-one rationales. By normalizing sufficiency and comprehensiveness scores against this value, we can estimate how faithful rationales are relative to the baseline fidelity of the model.

We use min-max normalization to normalize sufficiency and comprehensiveness with this null dif-



(a) Null difference. (b) Normalized sufficiency. (c) Normalized comprehensiveness.

Figure 4: Normalization is critical for interpreting sufficiency and comprehensiveness. Here we show evaluations of human-generated rationales based on RoBERTa.

ference. Formally, we define the metrics as follows:

$$\text{NullDiff}(\boldsymbol{x}, \hat{y}) = \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \mathbf{0})) \quad (3)$$

$$\text{NormSuff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \frac{\text{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0})}{1 - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0})} \quad (4)$$

$$\text{NormComp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \frac{\text{Comp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha})}{\text{Comp}(\boldsymbol{x}, \hat{y}, \mathbf{1})} \quad (5)$$

where $\hat{y} = \arg\max_y p(y|\boldsymbol{x})$. Note that $\text{NullDiff}(\boldsymbol{x}, \hat{y}) = 1 - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0}) = \text{Comp}(\boldsymbol{x}, \hat{y}, \mathbf{1})$. We clip NormSuff and NormComp between 0 and 1.

Fig. 4a shows the null difference for RoBERTa across all datasets by class. Significant variation exists between classes, especially for WikiAttack, FEVER, and E-SNLI, an observation that helps contextualize some of the results in Fig. 3, as reflected by the normalized fidelity metrics.

Fig. 4b shows that normalized sufficiency is much lower in the no-attack class in WikiAttack, meaning that no-attack rationales are barely more informative than an empty rationale. This resolves the puzzle that the short/empty rationales in the no-attack class have high sufficiency in Fig. 3b. It is also more consistent with the low comprehensiveness measured for these rationales.

Fig. 4c shows us that the comprehensiveness scores even out for FEVER under this normalization, suggesting that the previous result was simply a product of model bias. By contrast, the asymmetric scores for Movie and MultiRC shown in Fig. 3c cannot be explained by model bias, indicating that the interaction between task semantics and model learning may cause rationales to be more comprehensive in the negative class than in the positive class for these datasets.

Another outcome of normalization is to map sufficiency and comprehensiveness to the same scale. Comprehensiveness in single-text classification tasks are generally lower than that in document/query-style tasks.

These results suggest that sufficiency and comprehensiveness metrics are highly model-dependent and should not be compared across models without care.

Fidelity and model training. Examining how human rationale fidelity changes from epoch to epoch as models train (Fig. 5) further demonstrates the model-dependence of these measures. Random noise causes the models to have nonzero (but low) fidelity scores at epoch 0. However, we observe that even after accuracy stabilizes, sufficiency and comprehensiveness may continue to fluctuate significantly, e.g., FEVER sufficiency.[2] Further, the maximum fidelity may not co-occur with the maximum accuracy (e.g., MultiRC comprehensiveness). While most of the fluctuation isn't drastic, these differences could prove decisive in a head-to-head comparison of fidelity scores across different models or rationalization techniques. These observations suggest that we need to be cautious before claiming definitive fidelity for a given model using these automatic metrics.

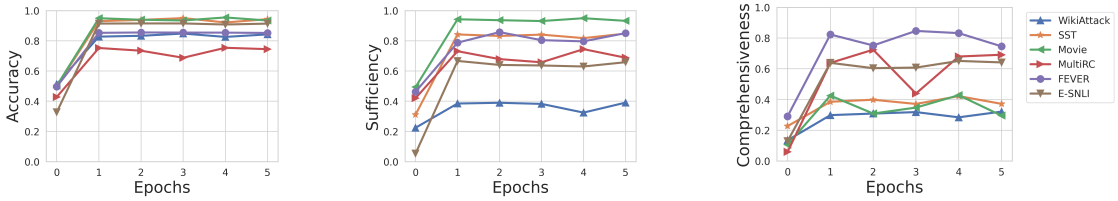
5 Characterizing Human Rationales beyond Sufficiency/Comprehensiveness

Sufficiency and comprehensiveness offer a limited perspective on the qualities of rationales. For example, does the 0.77 E-SNLI sufficiency reported in Fig. 2b correspond with a similar drop in accuracy, or do the rationales render the model less confident but equally accurate? And how can we distinguish between a highly concise rationale and one bloated with unnecessary information? We propose extensions of the basic fidelity framework to address these more nuanced questions.

5.1 Accuracy Evaluation with Rationales

Existing fidelity metrics measure differences in output probability rather than model performance, prompting the question of what is the practical effect of rationale fidelity. Moreover, they generally involve a model trained on complete texts but then evaluated on reduced texts based on rationales, ren-

[2] We observe similar issues with logistic regression, random forest, and LSTM. See the appendix.



(a) Accuracy (b) Normalized sufficiency (c) Normalized comprehensiveness

Figure 5: Accuracy, normalized sufficiency, and normalized comprehensiveness vs. #epochs in RoBERTa. While accuracy stabilizes after 1 epoch, sufficiency and comprehensiveness demonstrate significant fluctuation.

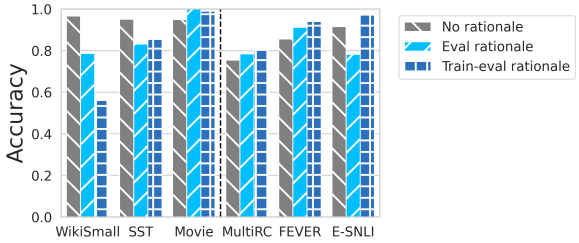


Figure 6: RoBERTa accuracy depending on whether we adapt models to rationale-only data. Human rationales are effective in improving accuracy in Movie and E-SNLI, but not in WikiAttack and SST.

	Training	Testing
No-rationale	No	No
Eval-rationale	No	Yes
Train-eval-rationale	Yes	Yes

Table 3: Use of rationales in different accuracy evaluations. The full-text model uses *no* rationale in training.

dering it unclear what outcome differences we can attribute to the missing information, and what to domain transfer between full and reduced text.

To answer these questions, we compare the accuracy of three variant training/evaluation regimes: 1) trained and evaluated on full text (*No rationale*); 2) trained on full text and evaluated on rationale-only text (*Eval rationale*); and 3) trained and evaluated on rationale-only text (*Train-eval rationale*).

The first variant is standard RoBERTa model training and evaluation. The second variant is the typical rationale evaluation setting: trained on full data and evaluated on reduced data. The third variant seeks to assess what performance gains can arise from model adaptation to the reduced data distribution. Table 3 summarizes the variants.[3]

Comparing the performance of these three models pits the benefits of data completeness (training on full information) against those of in-domain training (training on the same distribution as the evaluation data). If the former proves more valuable we would expect *Eval rationale* to outperform *Train-eval rationale*, and vice versa. In either case, we expect *No rationale* to have the best performance as it benefits from both qualities.

Fig. 6 shows some surprising divergences from these expectations. In four out of six cases, either *Train-eval rationale* accuracy or *Eval rationale* accuracy outperforms *No rationale* accuracy.

The effect of rationales in evaluation gives yet another perspective on the basic fidelity results presented in Fig. 2b. While the 0.77 sufficiency for E-SNLI corresponds with a significant accuracy drop between *No rationale* and *Eval rationale*, the 0.85 sufficiency for MultiRC corresponds with an *increase* in accuracy across these variants. The almost identical sufficiency of SST corresponds with a drop. "Insufficient" explanations can improve model performance, which suggests caution in using fidelity based on output probability as the sole arbiter of explanation quality.

The effect of model adaptation has interesting implications as well. We observe an improvement in performance from *Eval rationale* to *Train-eval rationale* in 4 out of 6 datasets, significant in the case of E-SNLI. In 3 out of 4 of these cases, the performance of *Train-eval rationale* also exceeds that of the *No rationale* setting.

This result is a hopeful sign for the topic area of *learning-from-explanation*, which seeks to use explanations as additional training supervision for models (Hancock et al., 2018; Zaidan and Eisner, 2008). It tells us that for a majority of our datasets, a perfectly human-mimicking rationale layer could boost the accuracy of a model's predictions. It is even possible that a version of this analysis could be used as a preliminary assessment of the usefulness of a rationale dataset as accuracy-boosting signal, though we leave this for future work.

In summary, from a model accuracy perspective, the quality of human rationales is strong for FEVER, MultiRC, and Movie, mixed for E-SNLI, and poor for SST and WikiAttack. This provides a somewhat different view from Fig. 4b. For exam-

[3]We only have human rationales on 1,049 instances in WikiAttack, so we use a different train/dev/test split from §3.

	Sufficiency	Comprehensiveness
brevity	fast drop	fast drop
redundancy	slow drop	fast drop
irrelevance	slow drop	slow drop
dependency	fast drop	slow drop

Table 4: Implications of irrelevance and redundancy on sufficiency and comprehensiveness.

ple, human rationales in MultiRC has lower (normalized) sufficiency based on output probability than SST but provide better accuracy sufficiency.

5.2 Fidelity Curves

Sufficiency and comprehensiveness struggle to convey more fine-grained qualities of human rationales. One problem that is not revealed by these measures is irrelevance. A rationale can be crammed with tokens that are not pertinent to prediction and still have high sufficiency and comprehensiveness, the most extreme example being a rationale that comprises the entire text.

We propose to assess rationale irrelevancy by looking at how sufficiency and comprehensiveness degrade as tokens are removed from the rationale. A rationale bloated with many irrelevant tokens should demonstrate a slow dropoff in sufficiency as tokens are removed, since many of these tokens will not be contributory. A rationale with more informational brevity should show a faster drop, as tokens are removed which were needed for prediction. We assess this by creating a "sufficiency curve" which traces this degradation at higher and higher occlusion rates.

In general, we suggest that a slow drop in sufficiency can be attributed to irrelevant or redundant tokens, while a fast drop in sufficiency can be due to dropping tokens that are either individually predictive or pieces of dependencies where multiple tokens are required to make a prediction. We can tell the difference by looking at the comprehensiveness curve — if individually predictive tokens are leaked into the rationale complement, the comprehensiveness should fall quickly, while if pieces of dependencies are, it should fall slowly. Table 4 summarizes our expectations.

We construct these fidelity curves as follows: For a given rationale $\boldsymbol{\alpha}$ and each of a series of replacement rates $R = 0, 0.05, 0.1, ..., 1.0$, we generate a reduced mask $\boldsymbol{\alpha}_r$ by randomly setting r fraction of tokens to 0 from the rationale. By calculating the mean normalized sufficiency and comprehensiveness over several trials for each replacement rate, we can draw a "sufficiency curve" (Fig. 7a) and a "comprehensiveness curve" (Fig. 7b).

Movie, WikiAttack, and SST exhibit slow drops in their sufficiency curves, showing that rationales in these datasets contain relatively many irrelevant or redundant tokens, and therefore remain sufficient even as some of their tokens are removed. Their comprehensiveness curves complete the story. The curves for all three datasets show relatively fast drops, implying redundancy rather than irrelevancy.

In comparison, E-SNLI, FEVER, and MultiRC all display relatively fast drops in sufficiency, implying fewer irrelevant or redundant tokens. They demonstrate generally higher comprehensiveness but somewhat different shapes (E-SNLI and MultiRC mostly show a slow drop, indicating dependence, while FEVER shows a fast drop, indicating irrelevance). The difference here between FEVER and MultiRC is interesting as they are similar in task, text, and rationale properties (Table 2). A possible explanation is that rationales in MultiRC are designed to consist of multiple mutually-dependent sentences whereas those of FEVER are single contiguous snippets of the text. This greater level of dependency is thus reflected in the slow-dropping comprehensiveness curve of MultiRC.

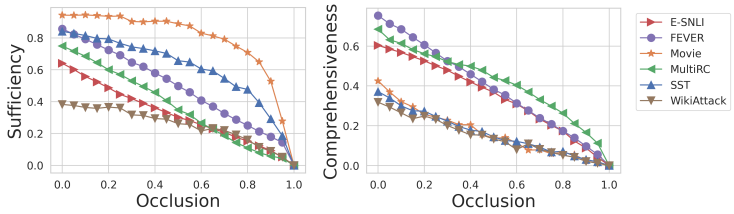
Hence, we find that human rationales for the three classification tasks are characterized by redundancy in human rationales, particularly Movie. The three document/query-style datasets, by contrast, are characterized by a relatively high degree of token dependency, explaining their relatively high comprehensiveness in Fig. 4c. While this observation is intuitive given the semantics of these tasks, it demonstrates the effectiveness of the proposed fidelity curves.

6 Related Work

We summarize additional related work in the following three areas.

Feature attribution. Feature attribution seeks to explain model behavior by attributing model predictions to specific inputs. Popular techniques include LIME (Ribeiro et al., 2016), integrated gradients (Sundararajan et al., 2017), SHAP (Lundberg and Lee, 2017), and attention mechanisms (Lei et al., 2016; Paranjape et al., 2020).

Human rationales. Many recent datasets in NLP have been released with rationales accompanying the document-level labels. ERASER (DeYoung et al., 2020) includes three additional datasets: CoS-



(a) Sufficiency (b) Comprehensiveness

	Sufficiency	Comprehensiveness
WikiAttack	slow drop	fast drop
SST	slow drop	fast drop
Movie	slow drop	fast drop
MultiRC	fast drop	fast→slow drop
FEVER	fast drop	fast drop
E-SNLI	fast drop	slow drop

(c) Summary by dataset.

Figure 7: Fidelity curves for all datasets (normalized sufficiency and comprehensiveness). Human rationales tend to be redundant in the single-text classification datasets, and dependent in the document/query-style datasets.

E (Rajani et al., 2019), BoolQ (Clark et al., 2019), and Evidence Inference (Lehman et al., 2019). Other rationale datasets include that of Kaushik et al. (2019) and Sen et al. (2020).

Attribution evaluation. A growing amount of work seeks to evaluate the quality of feature attribution. Beyond collecting human rationales as a gold-standard, a common human-based method is to test the utility of attribution masks in task-based human subject experiments (Carton et al., 2020; Lai and Tan, 2019; Poursabzi-Sangdeh et al., 2018; Lage et al., 2018; Lai et al., 2020).

Automatic model-based metrics beyond sufficiency and comprehensiveness include local model fidelity (Ribeiro et al., 2016), switching point (Nguyen, 2018), and area-over-the-perturbation-curve (Samek et al., 2016).

7 Concluding Discussion

Human explanations contain a lot of promise. The explainable AI community hopes to use them as a guide for evaluating model explanations and, possibly, for teaching models to make robust and well-reasoned decisions. In this work, we contribute to that effort by analyzing human rationales through the lens of automatic rationale evaluation methods, namely, sufficiency and comprehensiveness. We find that human rationales do not necessarily have high sufficiency or comprehensiveness.

Interpreting fidelity variance. Furthermore, there exists significant variance across datasets and classes. In §5.2, we speculate that some of these differences (e.g., dependency) can be explained by the semantic differences between classification and document/query-style tasks.

However, with such a small sample size of datasets ($n = 6$), it is difficult to determine whether these differences are due solely to task type or to other factors such as annotation instructions or individual dataset semantics. WikiAttack and E-SNLI, for example, display class asymmetry in their rationales, which likely contribute to their outlier status in Fig. 4 and 6 respectively. As we note in Fig. 2, modeling outcomes also have a heavy impact on explanation fidelity. While E-SNLI comprises an even class balance, our model learns a strong bias in favor of the neutral class, which contributes to a class imbalance in fidelity for that dataset (Fig. 4).

As more human-rationale datasets are released, it will become increasingly possible to categorize them by rationale properties. Our goal is to highlight the variance in these properties and call for more widespread empirical evaluations thereof.

Actionable implications. When human rationales are found to be unfaithful, this can mean that either they fail to capture relevant signal, or that the model improperly utilizes that signal, perhaps as a result of learning spurious associations. In either case, analysis can expose inconsistencies between human and model understanding of the task.

We propose three ways to extend fidelity metrics: normalization, model adaptation, and random ablation. Each addresses one shortcoming of the basic metric: normalization addresses the differences in class biases across models, adaptation the problem of domain inconsistency between full and rationale-only data, and ablation the inability of existing metrics to capture qualities like redundancy. While not all of these issues are salient for every application involving rationale fidelity, we offer them as potential solutions where necessary.

Overall, our results suggest that the idea of one-size-fits-all fidelity benchmarks might be problematic: human rationales may not be simply treated as gold standard. We need to design careful procedures to collect human rationales, understand properties of the resulting human rationales, and cautiously interpret the evaluation metrics.

Acknowledgments

We thank helpful comments from anonymous reviewers. This work was supported in part by research awards from Amazon and Salesforce, and NSF IIS-1927322, 1941973.

References

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 EMNLP*.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Attention-Based Explanations Don't Help Humans Detect Misclassifications of Online Toxicity. In *Proceedings of ICWSM*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL*.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of ICLR*.

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)*.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. An Evaluation of the Human-Interpretability of Explanation. In *Proceedings of NeurIPS*.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of CHI*.

Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of FAT**.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of NAACL*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of NeurIPS*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of NAACL*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of EMNLP*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proceedings of NeurIPS*.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. *arXiv preprint*. ArXiv: 1802.07810.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of KDD*.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of ACL*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of ACL*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of ICML*.

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

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of EMNLP*.

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]*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of WWW*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of EMNLP*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of EMNLP*.

A Derivation of Rationales for SST

The Stanford Sentiment Treebank (SST) consists of 9,620 short movie review snippets formatted as syntactic trees with a sentiment label in [-2,2] for each node, ranging from the single-token leaf nodes to the top-level node corresponding to the whole snippet.

We use a heuristic algorithm for flattening this representation into a 1-dimensional rationale for each document: beginning with the top node and traversing the tree in a breadth-first manner, we consider a node to be part of the rationale if the magnitude of its sentiment is greater than that of any of its descendants. That is, if the sentiment of a node cannot be explained by any of its syntactic constituents, then we consider it to be explanatory and include it in the top-level rationale.

Practically speaking, this results in a rationale dataset that is comprehensive by design, including all high-sentiment words and phrases that could explain the overall sentiment of each snippet. Table 5 shows a few examples of the resultant rationales.

B Model Implementation Details

We consider the following models:

- Logistic regression. We use the scikit-Learn implementation of logistic regression (Pedregosa et al., 2011), scanning across regularization constant ($C = \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$).
- Random forest. We use the scikit-Learn implementation of random forests, scanning across number of estimators ($\{16, 32, 64, 128, 256, 512\}$).
- LSTM (Hochreiter and Schmidhuber, 1997). We use the Pytorch (Paszke et al., 2017) implementation of a 1-layer BiLSTM, tuning across hidden layer size ($\{100, 200, 300\}$) and learning rate ($\{5e^{-4}, 1e^{-3}, 2e^{-3}\}$).
- RoBERTA (Liu et al., 2019). We use the HuggingFace (Wolf et al., 2020) pretrained distribution of this model with roughly 117m parameters. We tune the learning rate across values $\{5e^{-6}, 1e^{-5}, 2e^{-5}\}$, with 50 linear warmup steps.

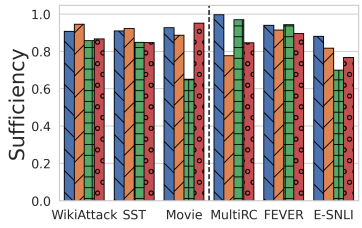
We train all LSTM models for 10 epochs and RoBERTa models for 5 epochs, tuning on development set accuracy. All neural network training was done on two 24G Nvidia Titan RTX GPUs. Training time varied from dataset to dataset, from minutes for SST to roughly 6 hours per model for E-SNLI.

To apply masking, we simply remove the tokens corresponding with 0s in the rationale mask. We always keep special tokens such as [CLS] and [SEP].

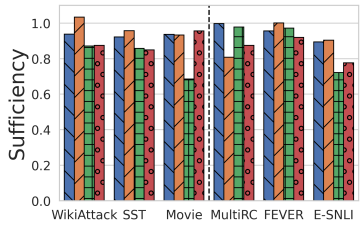
Following DeYoung et al. (2020), we flatten the three document/query-style datasets to single documents by simply appending the query to the document with a "[SEP]" token.

Rationale	Class
All the performances are **top notch** and , once you get through the accents , All or Nothing becomes an emotional , though **still positive ,** wrench of a sit .	Pos
While **surprisingly sincere** , this average little story is adorned with some **awesome** action photography and surfing	Pos
A **dreary rip-off** of **Goodfellas** that serves as a **muddled and offensive cautionary tale** for Hispanic **Americans**	Neg
A **long-winded and stagy** session of romantic contrivances **that never** really gels like the **shrewd feminist fairy tale** it could have been	Neg

Table 5: Example SST rationales generated by heuristic flattening procedure.

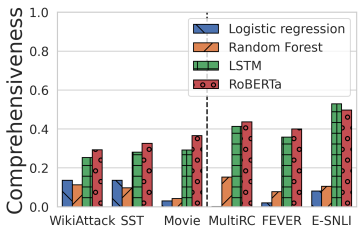


(a) Clipped sufficiency

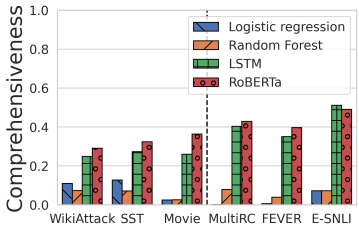


(b) ERASER sufficiency

Figure 8: Clipped sufficiency vs. ERASER sufficiency.



(a) Clipped comprehensiveness



(b) ERASER comprehensiveness

Figure 9: Clipped comprehensiveness vs. ERASER comprehensiveness.

C Eraser Sufficiency/Comprehensiveness vs. Our Definitions

Our definition of sufficiency and comprehensiveness diverge from that of DeYoung et al. (2020) in clipping the absolute difference between the full and rationalized class probability. This choice erases negative probability differences, cases where the rationalization makes the predicted class more probable than it already was. We do this as a way to bound fidelity metrics between 0 and 1. It also serves to simplify the mathematics of normalization, but practically speaking we find that it makes little difference (Fig. 8 and Fig. 9).

D The Effect of Normalization

We discuss the effect of normalization by class in §4. Fig. 10 compares the non-normalized against the normalized fidelity at the dataset level. This view makes clear the comprehensiveness gap between the classification datasets and the document/query-style datasets, and shows a wider range of sufficiency scores among the six datasets, when accounting for model bias.

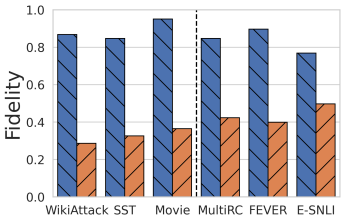
Fig. 11 shows the effect of normalization on fidelity scores for all models. We can see that it corrects the trend of weaker models showing better sufficiency that we observe in Fig. 2b, though logistic regression shows very high sufficiency and comprehensiveness for SST. Upon investigation, we find that this is because this model tends to have low confidence, often producing class probabilities between 0.5 and 0.7. This situation leads to relatively small null differences (Fig. 11a), which leads to the high observed comprehensiveness. In comparison, the null difference is substantially greater in deep models.

E The Effect of Hyperparameters and Training

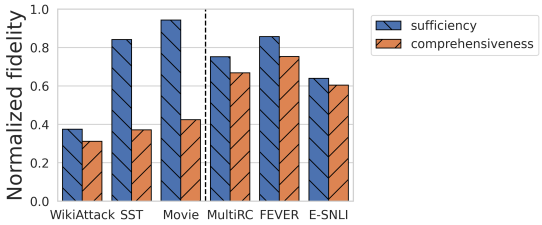
We largely focus on RoBERTa in this study because it is close the current state-of-the-art for NLP. However, we do some additional analysis on the other three models.

Fig. 12 shows how accuracy and rationale fidelity change with the value of the C regularization hyperparameter for the logistic regression model. Both the normalized sufficiency and comprehensiveness rise with model accuracy. The outlier is MultiRC, which is unable to achieve nontrivial accuracy, but which nevertheless experiences a rise in rationale fidelity.

The trends are less clear in Fig. 13, which tracks accuracy and fidelity over a range of numbers of



(a) Non-normalized fidelity



(b) Normalized fidelity

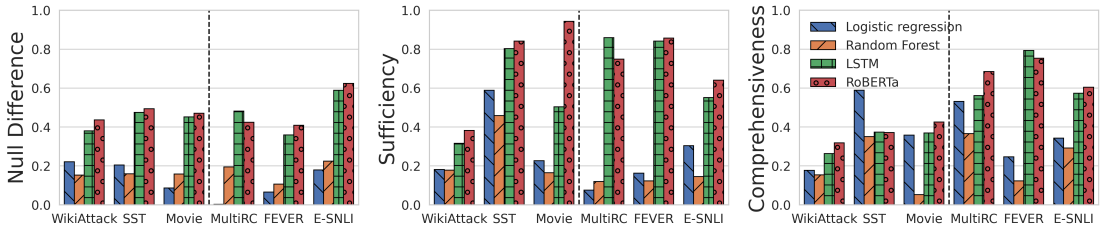
Figure 10: Non-normalized versus normalized fidelity.

estimators for the model. This may be because the accuracy of these models does not improve much with the increase in estimators.

Finally, Fig. 14 shows the change in accuracy and fidelity over training epochs for the LSTM model. We again see that fidelity metrics have a tendency to fluctuate when accuracy has seemingly stabilized, such as FEVER.

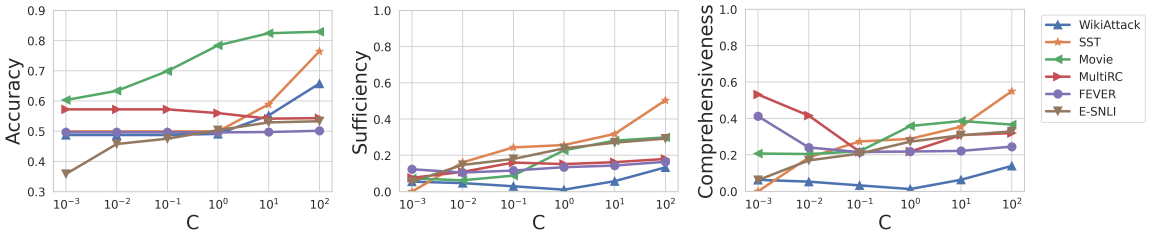
F Distribution of Fidelity Scores

Fig. 15 shows box plots of normalized fidelity scores for the six datasets. We see a wide range of variances. WikiAttack and E-SNLI, the two datasets with assymmetric rationales, display the highest variance in sufficiency, while WikiAttack, Movie, and MultiRC who relatively high variance in their comprehensiveness scores.



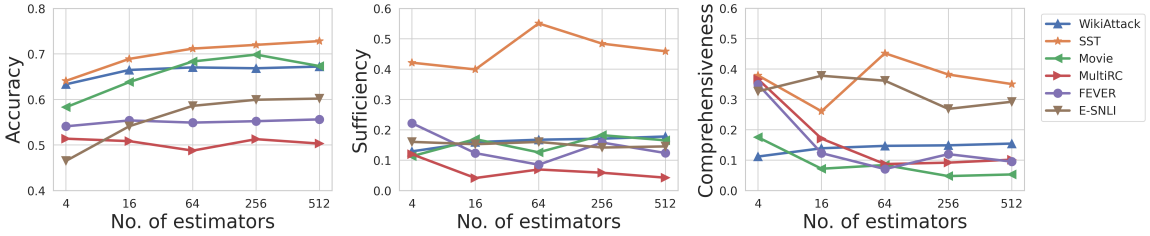
(a) Null difference (b) Normalized sufficiency (c) Normalized comprehensiveness

Figure 11: Normalized fidelity for all models.



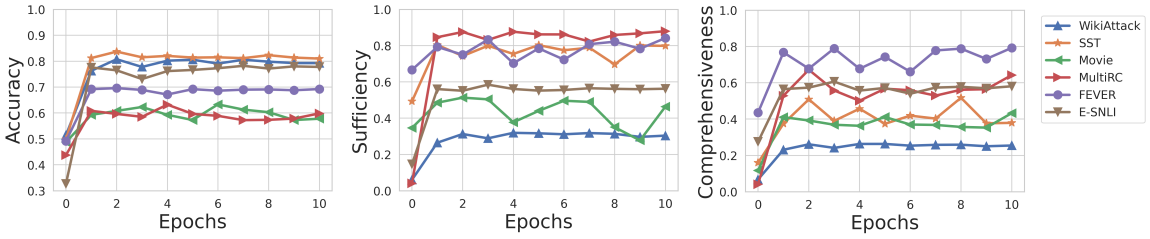
(a) Accuracy (b) Normalized sufficiency (c) Normalized comprehensiveness

Figure 12: Accuracy, sufficiency and comprehensiveness of logistic regressions models by regularization term C.



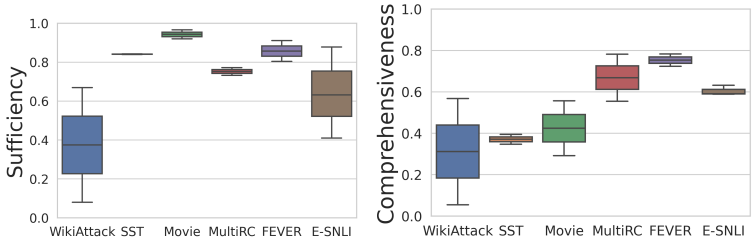
(a) Accuracy (b) Normalized sufficiency (c) Normalized comprehensiveness

Figure 13: Accuracy, sufficiency and comprehensiveness of random forest models by number of estimators.



(a) Accuracy (b) Normalized sufficiency (c) Normalized comprehensiveness

Figure 14: Accuracy, sufficiency and comprehensiveness of LSTM models by training epoch.



(a) Normalized sufficiency (b) Normalized comprehensiveness

Figure 15: Box plots of normalized fidelity metrics.