# Can Rationalization Improve Robustness?

**Howard Chen** **Jacqueline He** **Karthik Narasimhan** **Danqi Chen**
Department of Computer Science, Princeton University
{howardchen, karthikn, danqic}@cs.princeton.edu
jyh@princeton.edu

## Abstract

A growing line of work has investigated the development of neural NLP models that can produce *rationales*—subsets of input that can explain their model predictions. In this paper, we ask whether such rationale models can provide robustness to adversarial attacks in addition to their interpretable nature. Since these models need to first generate rationales ("rationalizer") before making predictions ("predictor"), they have the potential to ignore noise or adversarially added text by simply masking it out of the generated rationale. To this end, we systematically generate various types of 'AddText' attacks for both token and sentence-level rationalization tasks and perform an extensive empirical evaluation of state-of-the-art rationale models across five different tasks. Our experiments reveal that rationale models show promise in improving robustness but struggle in certain scenarios—e.g., when the rationalizer is sensitive to position bias or lexical choices of the attack text. Further, leveraging human rationales as supervision does not always translate to better performance. Our study is a first step towards exploring the interplay between interpretability and robustness in the rationalize-then-predict framework.[1]

## 1 Introduction

Rationale models aim to introduce a degree of interpretability into neural networks by implicitly baking in explanations for their decisions (Lei et al., 2016; Bastings et al., 2019; Jain et al., 2020). These models are carried out in a two-stage 'rationalize-then-predict' framework, where the model first selects a subset of the input as a *rationale* and then makes its final prediction for the task solely using the rationale. A human can then inspect the selected rationale to verify the model's reasoning over the most relevant parts of the input for the prediction at hand.

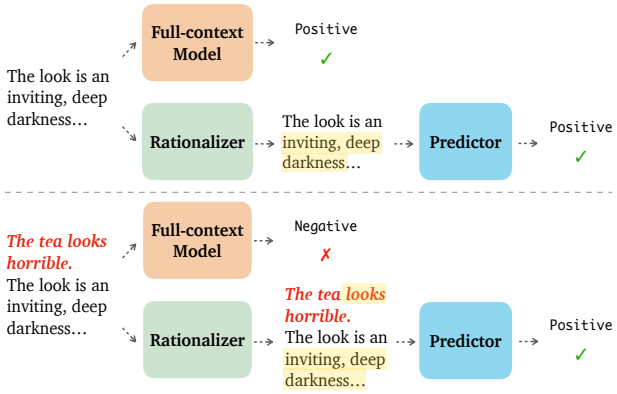


Figure 1: Top: input text is processed by a rationale model (rationalizer and predictor) and a full-context model (making predictions based on the whole input) separately in a *beer review* sentiment classification dataset. Both models make correct predictions. Bottom: when an attack sentence "*The tea looks horrible.*" is inserted, the full-context model fails. The rationalizer successfully excludes the negative-sentiment word "horrible" from the selected rationales (yellow highlights). The predictor is hence not distracted by the attack sentence.

While previous work has mostly focused on the plausibility of extracted rationales and whether they represent faithful explanations (DeYoung et al., 2020), we ask the question of how rationale models behave under adversarial attacks (i.e., do they still provide plausible rationales?) and whether they can help improve robustness (i.e., do they provide better task performance?). Our motivation is that the two-stage decision-making could help models ignore noisy or adversarially added text within the input. For example, Figure 1 shows a state-of-the-art rationale model (Paranjape et al., 2020) smoothly handles input with adversarially added text by selectively masking it out during the rationalization step. Factorizing the rationale prediction from the task itself effectively 'shields' the predictor from

[1]Our code is publicly available at: https://github.com/princeton-nlp/rationale-robustness.

having to deal with adversarial inputs.

To answer these questions, we first generate adversarial tests for a variety of popular NLP tasks (§4). We focus specifically on model-independent, 'AddText' attacks (Jia and Liang, 2017), which augment input instances with noisy or adversarial text at *test time*, and study how the attacks affect rationale models both in their prediction of rationales and final answers. For diversity, we consider inserting the attack sentence at different positions of context, as well as three types of attacks: random sequences of words, arbitrary sentences from Wikipedia, and adversarially-crafted sentences.

We then perform an extensive empirical evaluation of multiple state-of-the-art rationale models (Paranjape et al., 2020; Guerreiro and Martins, 2021), across five different tasks that span review classification, fact verification, and question answering (§5). In addition to the attack's impact on task performance, we also assess rationale prediction by defining metrics on gold rationale coverage and attack capture rate. We then investigate the effect of incorporating human rationales as supervision, the importance of attack positions, and the lexical choices of attack text. Finally, we investigate an idea of improving the rationalizer by adding augmented pseudo-rationales during training (§7).

Our key findings are the following:

1. Rationale models show promise in providing robustness. Under our strongest type of attack, rationale models in many cases achieve less than 10% drop in task performance while full-context models suffer more (11%–27%).
2. However, robustness of rationale models can vary considerably with the choice of lexical inputs for the attack and is quite sensitive to the attack position.
3. Training models with explicit rationale supervision does not guarantee better robustness to attacks. In fact, their accuracy drops under attack are higher by 4-10 points compared to rationale models without supervision.
4. Performance under attacks is significantly improved if the rationalizer can effectively mask out the attack text. Hence, our simple augmented-rationale training strategy can effectively improve robustness (up to 4.9%).

Overall, our results indicate that while there is promise in leveraging rationale models to improve robustness, current models may not be sufficiently equipped to do so. Furthermore, adversarial tests may provide an alternative form to evaluate rationale models in addition to prevalent plausability metrics that measure agreement with human rationales. We hope our findings can inform the development of better methods for rationale predictions and instigate more research into the interplay between interpretability and robustness.

## 2 Related Work

**Rationalization** There has been a surge of work on explaining predictions of neural NLP systems, from post-hoc explanation methods (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017), to analysis of attention mechanisms (Jain and Wallace, 2019; Serrano and Smith, 2019). We focus on *extractive rationalization* (Lei et al., 2016), which generates a subset of inputs or highlights as "rationales" such that the model can condition predictions on them. Recent development has been focusing on improving joint training of rationalizer and predictor components (Bastings et al., 2019; Yu et al., 2019; Jain et al., 2020; Paranjape et al., 2020; Guerreiro and Martins, 2021; Sha et al., 2021), or extensions to text matching (Swanson et al., 2020) and sequence generation (Vafa et al., 2021). These rationale models are mainly compared based on predictive performance, as well as agreement with human annotations (DeYoung et al., 2020). In this work, we question how rationale models behave under adversarial attacks and whether they can provide robustness benefits through rationalization.

**Adversarial examples in NLP** Adversarial examples have been designed to reveal the brittleness of state-of-the-art NLP models. A flood of research has been proposed to generate different adversarial attacks (Jia and Liang, 2017; Iyyer et al., 2018; Belinkov and Bisk, 2018; Ebrahimi et al., 2018, *inter alia*), which can be broadly categorized by types of input perturbations (sentence-, word- or character-level attacks), and access of model information (black-box or white-box). In this work, we focus on *model-independent*, label-preserving attacks, in which we *insert* a random or an adversarially-crafted sentence into input examples (Jia and Liang, 2017). We hypothesize that a good extractive rationale model is expected to learn to ignore these distractor sentences and hence achieve better performance under attacks.

**Interpretability and robustness** A key motivation of our work is to bridge the connection be-

tween interpretability and robustness, which we believe is an important and under-explored area. Alvarez-Melis and Jaakkola (2018) argue that robustness of explanations is a key desideratum for interpretability. Slack et al. (2020) explore unreliability of attribution methods against input perturbations. Camburu et al. (2020) introduce an adversarial framework to sanity check models against their generated inconsistent free-text explanations. Zhou et al. (2020) propose to evaluate attribution methods through dataset modification. Noack et al. (2021) show that image recognition models can achieve better adversarial robustness when they are trained to have interpretable gradients. To the best of our knowledge, we are the first to quantify the performance of rationale models under textual adversarial attacks and understand whether rationalization can inherently provide robustness.

## 3 Background

Extractive rationale models[2] output predictions through a two-stage process: the first stage ("rationalizer") selects a subset of the input as a *rationale*, while the second stage ("predictor") produces the prediction using only the rationale as input. *Rationales* can be any subset of the input, and we characterize them roughly into either token-level or sentence-level rationales, which we will both investigate in this work. The task of predicting rationales is often framed as a binary classification problem over each atomic unit depending on the type of rationales. The rationalizer and the predictor are often trained jointly using task supervision, with gradients back-propagated through both stages. We can also provide explicit rationale supervision, if human annotations are available.

### 3.1 Formulation

Formally, let us assume a supervised classification dataset $\mathcal{D} = \{(x, y)\}$ , where each input $x = x_1, x_2, ..., x_T$ is a concatenation of $T$ sentences and each sentence $x_t = (x_{t,1}, x_{t,2}, ...x_{t,n_t})$ contains $n_t$ tokens, and $y$ refers to the task label. A rationale model consists of two main components: 1) a rationalizer module $z = R(x; \theta)$, which generates a discrete mask $z \in \{0, 1\}^L$ such that $z \odot x$ selects a subset from the input ($L = T$ for sentence-level rationalization or $L = \sum_{i=1}^{T} n_i$ for token-level rationales), and 2) a predictor module $\hat{y} = C(x, z; \phi)$ that makes a prediction $\hat{y}$ using the generated rationale $z$. The entire model $M(x) = C(R(x))$ is trained end-to-end using the standard cross-entropy loss. We describe detailed training objectives in §5.

### 3.2 Evaluation

Rationale models are traditionally evaluated along two dimensions: a) their downstream task performance, and b) the quality of generated rationales. To evaluate rationale quality, prior work has used metrics like token-level F1 or Intersection Over Union (IOU) scores between the predicted rationale and a human rationale (DeYoung et al., 2020):

$$\text{IOU} = \frac{|z \cap z^*|}{|z \cup z^*|},$$

where $z^*$ is the human-annotated gold rationales.

A good rationale model should not sacrifice task performance while generating rationales that concur with human rationales. However, metrics like F1 score may not be the most appropriate way to capture this as it only captures *plausibility* instead of *faithfulness* (Jacovi and Goldberg, 2020).

## 4 Robustness Tests for Rationale Models

### 4.1 AddText Attacks

Our goal is to construct attacks that can test the capability of extractive rationale models to ignore spurious parts of the input. Broadly, we used two guiding criteria for selecting the type of attacks: 1) they should be additive since an extractive rationale model can only "ignore" the irrelevant context. For other attacks such as counterfactually edited data (CAD) (Kaushik et al., 2020), even if the rationalizer could identify the edited context, the predictor is not necessarily strong enough to reason about the counterfactual text; 2) they should be model-independent since our goal is to compare the performance across different types of rationale and baseline models. Choosing strong gradient-based attacks (Ebrahimi et al., 2018; Wallace et al., 2019) would probably break all models, but that is beyond the scope of our hypothesis. An attack is suitable as long as it reduces performance of standard classification models by a non-trivial amount (our attacks reduce performance by 10%–36% absolute).

Keeping these requirements in mind, we focus on label-preserving text addition attacks Jia and Liang (2017), which can test whether rationale

[2]Abstractive models (Wiegreffe et al., 2021; Narang et al., 2020), which generate rationales as free text, are an alternative class of models that we do not consider in this work.

models are invariant to the addition of extraneous information and remain consistent with their predictions. Attacks are only added at test time and are not available during model training.

**Attack construction** Formally, an AddText attack $A(x)$ modifies the input $x$ by adding an attack sentence $x_{\text{adv}}$, without changing the ground truth label $y$. In other words, we create new perturbed test instances $(A(x), y)$ for the model to be evaluated on. While some prior work has considered the addition of a few tokens to the input (Wallace et al., 2019), we add complete sentences to each input, similar to the attacks in Jia and Liang (2017). This prevents unnatural modifications to the existing sentences in the original input $x$ and also allows us to test both token-level and sentence-level rationale models (§5.1). We experiment with adding the attack sentence $x_{\text{adv}}$ either at the beginning or the end of the input $x$.[3]

**Types of attacks** We explore three different types of attacks: (1) **AddText-Rand**: we simply add a random sequence of tokens uniformly sampled from the task vocabulary. This is a weak attack that is easy for humans to spot and ignore since it does not guarantee grammaticality or fluency. (2) **AddText-Wiki**: we add an arbitrarily sampled sentence from English Wikipedia into the task input (e.g., "Sonic the Hedgehog, designed for ..."). This attack is more grammatical than AddText-Rand, but still adds text that is likely irrelevant in the context of the input $x$. (3) **AddText-Adv**: we add an adversarially constructed sentence that has significant lexical overlap with tokens in the input $x$ while ensuring the output label is unchanged. This type of attack is inspired by prior attacks such as AddOneSent (Jia and Liang, 2017) and is the strongest attack we consider since it is more grammatical, fluent, and contextually relevant to the task. The construction of this attack is also specific to each task we consider, hence we provide examples listed in Table 1 and more details in §5.3.

### 4.2 Robustness Evaluation

We measure the robustness of rationale models under our attacks along two dimensions: *task performance*, and *generated rationales*. The change in task performance is simply computed as the difference between the average scores of the model on the original vs perturbed test sets:

$$\Delta = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} f(M(x), y) - f(M(A(x)), y),$$

where $f$ denotes a scoring function (F1 scores in extractive question answering and $\mathbb{I}(y = \hat{y})$ in text classification). To measure the effect of the attacks on rationale generation, we use two metrics:

**Gold rationale F1 (GR)** This is defined as the F1 score between the predicted rationale and a human-annotated rationale, either computed at the token or sentence level. The token-level GR score is equivalent to F1 scores reported in previous work (Lei et al., 2016; DeYoung et al., 2020). A good rationalizer should generate plausible rationales and be not affected by the addition of attack text.

**Attack capture rate (AR)** We define AR as the recall of the inserted attack text in the rationale generated by the model:

$$\text{AR} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\sim\mathcal{D}} \frac{|x_{\text{adv}} \cap (z \odot A(x))|}{|x_{\text{adv}}|},$$

where $x_{\text{adv}}$ is the attack sentence added to each instance (i.e., $A(x)$ is the result of inserting $x_{\text{adv}}$ into $x$), $z \odot A(x)$ is the predicted rationale. The metric above applies on both token or sentence level ($|x_{\text{adv}}| = 1$ for sentence-level rationalization and number of tokens in the attack sentence for token-level rationalization). This metric allows us to measure how often a rationale model can *ignore* the added attack text—a maximally robust rationale model should have an AR of 0.

## 5 Models and Tasks

We investigate two different state-of-the-art selective rationalization approaches: 1) sampling-based stochastic binary mask generation (Bastings et al., 2019; Paranjape et al., 2020), and 2) deterministic sparse attention through constrained inference (Guerreiro and Martins, 2021). We adapt these models, using two separate BERT encoders for the rationalizer and the predictor, and consider training scenarios with and without explicit rationale supervision. We also consider a full-context model as baseline. We provide a brief overview of each model here and leave details including loss functions and training to §A.1.

[3] In §6.4, we also consider inserting the attack sentence at a random position for studying the effect of attack positions.

| Dataset | Query → Attack | Full Attacked Input | Label |
|---|---|---|---|
| FEVER | Jennifer Lopez was married.<br>→ Jason Bourne was unmarried. | *Query*: Jennifer Lopez was married.<br>*Context*: Jennifer Lynn Lopez (born July 24 , 1969), also known as JLo, is an American singer . . . She subsequently married longtime friend Marc Anthony . . . Jason Bourne was unmarried. | Supports |
| SQuAD | Where did Super Bowl 50 take place?<br>→ The Champ Bowl 40 took place in Chicago. | *Query*: Where did Super Bowl 50 take place?<br>*Context*: Super Bowl 50 was an American football game to determine the champion . . . was played on February 7, 2016, at Levi's Stadium . . . The Champ Bowl 40 took place in Chicago. | Levi's Stadium |
| Beer | N/A<br>→ The tea looks horrible. | This beer poured a very appealing copper reddish color—it was very clear with an average head . . . The tea looks horrible. | Positive |

Table 1: AddText-Adv attack applied to three datasets. The query (blue) is transformed into an attack (red). The query together with the context forms the input. The attack is inserted to the context. We only show insertion at the end, but the attack can be inserted at any position between sentences. A model needs to associate the query and the evidence (ground truth rationale) in the context and not be distracted by the inserted attack to make the correct prediction. Note that the Beer dataset doesn't have a query and the attack sentence is dependent on the label (§5.3).

### 5.1 Models without Rationale Supervision

**Variational information bottleneck (VIB)** This model (Paranjape et al., 2020) imposes a discrete bottleneck objective (Alemi et al., 2017) to select a mask $z \in \{0,1\}^L$ from the input $x$. The rationalizer samples $z$ using Gumbel-Softmax and the predictor uses only $z \odot x$ for the final prediction. During inference, we select the top-$k$ scored rationales, where $k$ is determined by the sparsity $\pi$.

**Sparse structured text rationalization (SPECTRA)** This model (Guerreiro and Martins, 2021) extracts a deterministic structured mask $z$ by solving a constrained inference problem by applying factors to the global scoring function while optimizing the end task performance. The entire computation is deterministic and allows for back-propagation through the LP-SparseMAP solver (Niculae and Martins, 2020). We use the `BUDGET` factor to control the sparsity $\pi$.

**Full-context model (FC)** As a baseline, we also consider a full-context model, which makes predictions directly conditioned on the entire input. The model is a standard BERT model which adds task-specific classifiers on top of the encoder (Devlin et al., 2019). The model is trained with a cross-entropy loss using task supervision.

### 5.2 Models with Rationale Supervision

**VIB with human rationales (VIB-sup)** When human-annotated rationales $z^*$ are available, they can be used to guide the prediction of the sampled masks $z$ by adding a cross entropy loss between them (more details in §A.1). VIB-sup leverages this supervision signal to guide rationale prediction.

| Dataset | Rationale Granularity (w/ Human Rationale) | Task |
|---|---|---|
| FEVER | Sentence (✓) | Fact verification† |
| MultiRC | Sentence (✓) | Question answering† |
| SQuAD | Sentence (✓) | Question answering‡ |
| Beer | Token (✗) | Sentiment † |
| Hotel | Token (✗) | Sentiment † |

Table 2: Dataset characteristics for the five datasets. †: classification, ‡: span prediction tasks.

**Full-context model with human rationales (FC-sup)** We also extend the FC model to leverage human-annotated rationales supervision during training by adding a linear layer on top of the sentence/token representations. Essentially, it is multi-task learning of rationale prediction and the original task, shared with the same BERT encoder. The supervision is added by calculating the cross entropy loss between the human-annotated rationales and the predicted rationales (§A.1).

### 5.3 Tasks

We evaluate the models on five datasets that cover both sentence-level (FEVER, MultiRC, SQuAD) and token-level (Beer, Hotel) rationalization (examples in Table 1). We summarize the dataset characteristics in Table 2.

**FEVER** FEVER is a sentence-level binary classification fact verification dataset from the ERASER benchmark (DeYoung et al., 2020). The input contains a claim specifying a fact to verify and a passage of multiple sentences supporting or refuting the claim. For the AddText-Adv attacks, we add modified query text to the claims by replacing nouns and adjectives in the sentence with antonyms

from WordNet (Fellbaum, 1998).

**MultiRC** MultiRC (Khashabi et al., 2018) is a sentence-level multi-choice question answering task (reformulated as 'yes/no' questions). For the AddText-Adv attacks, we transform the question and the answer separately using the same procedure we used for FEVER.

**SQuAD** SQuAD (Rajpurkar et al., 2016) is a popular question answering dataset. We use the AddOneSent attacks proposed in Adversarial SQuAD (Jia and Liang, 2017), except that they always insert the sentence at the end of the paragraph and we consider inserting at the beginning, the end, and a random position. Since SQuAD does not contain human rationales, we use the sentence that contains the correct answer span as the ground truth rationale sentence. We report F1 score for SQuAD.

**Beer** BeerAdvocate is a multi-aspect sentiment analysis dataset (McAuley et al., 2012), modeled as a token-level rationalization task. We use the *appearance* aspect in out experiments. We convert the scores into the binary labels following Chang et al. (2020). This task does not have a query as in the previous tasks, we insert a sentence with the template "`{SUBJECT}` is `{ADJ}`" into a negative review where the adjective is positive (e.g., "The tea looks fabulous.") and vice versa. We consider one object "car" and eight adjectives (e.g., "clean/filthy", "new/old").

**Hotel** TripAdvisor Hotel Review is also a multi-aspect sentiment analysis dataset (Wang et al., 2010). We use the *cleanliness* aspect in our experiments. We generate AddText-Adv attacks in the same way as we did for the Beer dataset. We consider three objects ranging from more relevant words such as "tea" to less related word "carpet" and six adjectives (e.g., "pretty/disgusting", "good/bad", "beautiful/ugly").

## 6 Results

For all attacked test sets, we report the average scores with attack sentence inserted at the beginning and the end of the inputs. Our findings shed light on the relationship between GR, AR, and drop in performance, which eventually lead to a promising direction to improve performance of rationale models under attacks (§7).

### 6.1 Task Performance

Figure 2 summarizes the average scores on all datasets for each model under the three attacks we consider. We first observe that all models (including the full-context models FC and FC-sup) are mildly affected by AddText-Rand and AddText-Wiki, with score drops of around 1-2%. However, the AddText-Adv attack leads to more significant drops in performance for all models, as high as 46% for SPECTRA on the Hotel dataset. We break out the AddText-Adv results in a more fine-grained manner in Table 3. Our main observation is that the rationale models (VIB, SPECTRA, VIB-sup) are generally more robust than their non-rationale counterparts (FC, FC-sup) on four out of the five tasks, and in some cases dramatically better. For instance, on the Beer dataset, SPECTRA only suffers a 5.7% drop (95.4 $\rightarrow$ 89.7) compared to FC's huge 34.3% drop (93.8 $\rightarrow$ 59.5) under attack. The only exception is the Hotel dataset, where both the VIB and SPECTRA models perform worse under attack compared to FC. We analyze this phenomena and provide a potential reason below.

### 6.2 Robustness Evaluation: GR vs AR

In Table 4, we report the Gold Rationale F1 (GR) and Attack Capture Rate (AR) for all models. When attacks are added, GR consistently decreases for all tasks. However, AR ranges widely across datasets. VIB and SPECTRA have lower AR and higher GR compared to FC-sup across all tasks, which is correlated with their superior robustness to AddText-Adv attacks.

Next, we investigate the poor performance of VIB and SPECTRA on the Hotel dataset by analyzing the choice of words in the attack. Using the template "My car is `{ADJ}`.", we measure the percentage of times the rationalizer module selects the adjective as part of its rationale. When the adjectives are "dirty" and "clean", the VIB model selects them a massive 98.5% of the time. For "old" and "new", VIB still selects them 50% of the time. On the other hand, the VIB model trained on Beer reviews with attack template "The tea is `{ADJ}`." only selects the adjectives 20.5% of the time (when the adjectives are "horrible" and "fabulous"). This shows that the bad performance of the rationale models on the Hotel dataset is due to their inability to ignore task-related adjectives in the attack text, hinting that the lexical choices made in constructing the attack can largely impact robustness.

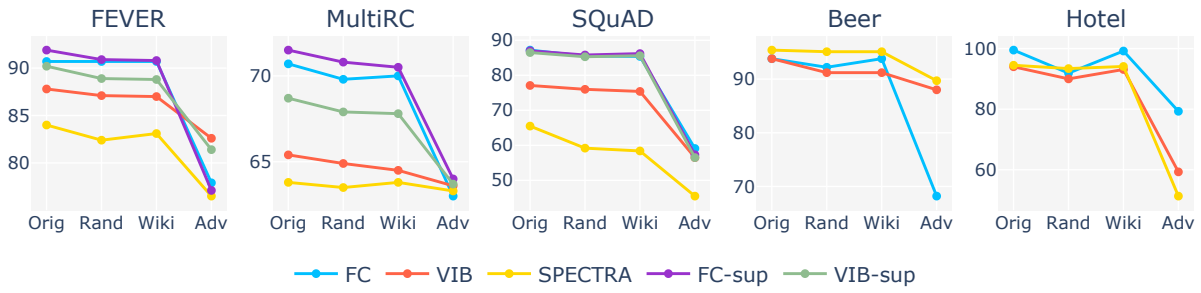


Figure 2: Original performance (Orig) and the three type of attacks AddText-Rand (Rand), AddText-Wiki (Wiki), and AddText-Adv (Adv) evaluated on five datasets and all of the models. FC-sup and VIB-sup models used human rationales during training (§5.2).

| | FEVER | | | MultiRC | | | SQuAD | | | Beer | | | Hotel | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ori | Att | Δ ↓ | Ori | Att | Δ ↓ | Ori | Att | Δ ↓ | Ori | Att | Δ ↓ | Ori | Att | Δ ↓ |
| Majority | 50.7 | - | - | 54.8 | - | - | - | - | - | 68.9 | - | - | 50.0 | - | - |
| FC | 90.7 | 77.9 | 12.8 | 70.7 | 63.0 | 7.7 | 87.2 | 59.1 | 28.1 | 93.8 | 59.5 | 34.3 | 99.5 | 79.3 | **20.2** |
| VIB | 87.8 | 82.6 | **5.2** | 65.4 | 63.6 | 1.8 | 77.1 | 56.5 | 20.6 | 93.8 | 88.0 | 5.8 | 94.0 | 59.3 | 34.8 |
| SPECTRA | 84.0 | 76.5 | 7.6 | 63.8 | 63.3 | **0.5** | 65.5 | 45.5 | **20.0** | 95.4 | 89.7 | **5.7** | 94.5 | 51.3 | 43.2 |
| FC-sup | 91.9 | 77.1 | 14.8 | 71.5 | 64.0 | 7.5 | 87.0 | 57.3 | **29.7** | - | - | - | - | - | - |
| VIB-sup | 90.2 | 81.4 | **8.8** | 68.7 | 63.7 | **5.0** | 86.5 | 56.5 | 30.0 | - | - | - | - | - | - |

Table 3: Original (Ori) versus attacked (Att) task performance on the five selected datasets under the AddText-Adv attack. We report accuracy for all datasets except for SQuAD, which we report F1. The attacked performance is the average of inserting the attack at the start and at the end of the text input.

| | FEVER | | MultiRC | | SQuAD | | Beer | | Hotel | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GR ↑ | AR ↓ | GR ↑ | AR ↓ | GR ↑ | AR ↓ | GR ↑ | AR ↓ | GR ↑ | AR ↓ |
| VIB | 36.9→30.3 | 59.4 | 15.8→13.9 | 35.8 | 86.2→84.9 | 63.7 | 20.5→18.1 | 11.9 | 23.5→22.6 | 18.4 |
| SPECTRA | 26.9→21.5 | 40.6 | 11.9→11.8 | 22.6 | 67.1→60.8 | 52.6 | 28.6→27.8 | 15.2 | 19.5→18.3 | 31.6 |
| FC-sup | 51.5→45.5 | 65.9 | 50.0→42.7 | 55.7 | 99.6→98.8 | 97.8 | - | - | - | - |
| VIB-sup | 50.6→44.3 | 67.0 | 36.1→22.7 | 58.7 | 99.5→97.8 | 97.2 | - | - | - | - |

Table 4: Gold rationale F1 (GR) (original→perturbed input) and attack capture rate (AR) for the AddText-Adv attack on the five tasks (defined in §4.2). The reported number is the average of inserting the attack at the start and at the end of the text input.

We examine where the rationale model gains robustness by inspecting the generated rationales. Table 5 shows the accuracy breakdown under attack for VIB and VIB-sup models. Intuitively, both models perform best when the gold rationale is selected and the attack is avoided, peaking at 91.1 for VIB and 92.4 for VIB-sup. Models perform much worse when the gold rationale is omitted and the attack is included (73.6 for VIB and 74.1 for VIB-sup), highlighting the importance of choosing good and skipping the bad as rationales.

### 6.3 Impact of Gold Rationale Supervision

Perhaps surprisingly, adding explicit rationale supervision does not help improve robustness (Table 3). Across FEVER, MultiRC and SQuAD, VIB-sup consistently has a higher Δ between its scores on the original and perturbed instances. We observe that models trained with human rationales generally have *higher GR*, but they also capture a *much higher AR* across the board (Table 4). On MultiRC, for instance, the VIB-sup model outperforms VIB in task performance because of its higher GR (36.1 versus 15.8). However, when under attack, VIB-sup's high 58.7 AR, hindering the performance compared to VIB, which has a smaller 35.8 AR. This highlights a potential shortcoming of prior work in only considering metrics like IOU (similar in spirit to GR) to assess rationale models. The finding also points to the risk of straightforwardly

| | VIB Acc (%) | VIB-sup Acc (%) |
|---|---|---|
| Original | 87.8 | 90.2 |
| Overall Attack | 83.0 (100%) | 84.9 (100%) |
| G ✓ A ✓ | 83.3 (34%) | 85.5 (77%) |
| G ✓ A ✗ | 91.1 (32%) | 92.4 (11%) |
| G ✗ A ✓ | 73.6 (22%) | 74.1 (12%) |
| G ✗ A ✗ | 77.7 (12%) | 68.0 (0%) |

Table 5: Accuracy breakdown of the VIB and VIB-sup models on the FEVER dataset. The attack is inserted at the beginning of the passage. ✓ indicates the Gold (G) or Attack (A) sentence is selected as rationale and ✗ otherwise. We show the percentage of examples in parenthesis. The highlighted row shows the desirable category and models achieve the highest accuracy.

incorporating supervised rationale as it could result in the existing model overfitting to them.

### 6.4 Sensitivity of Attack Positions

We further analyze the effect of attack text on rationale models by varying the attack position. Figure 3 displays the performance of VIB, VIB-sup and FC on FEVER and SQuAD when the attack sentence is inserted into the first, last or any random position in between. We observe performance drops on both datasets when inserting the attack sentence at the beginning of the context text as opposed to the end. For example, when the attack sentence is inserted at the beginning, the VIB model drops from 77.1 F1 to 40.9 F1, but it only drops from 77.1 F1 to 72.1 F1 for a last position attack on SQuAD. This hints that rationale models may implicitly be picking up positional biases from the dataset, similar to their full-context counterparts (Ko et al., 2020). We provide fine-grained plots for AR versus attack positions in §A.4.

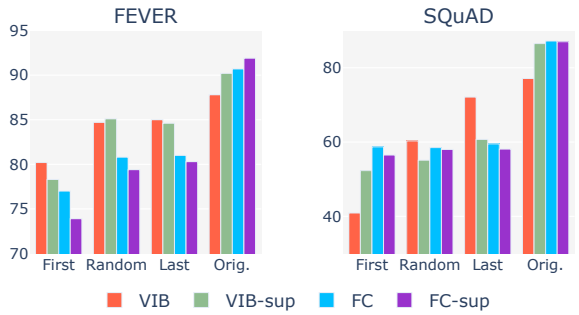


Figure 3: Accuracy when attack is inserted at different sentence positions, highlighting the positional bias picked up by the models.

| | FEVER Ori | FEVER Att | FEVER Δ ↓ | MultiRC Ori | MultiRC Att | MultiRC Δ ↓ |
|---|---|---|---|---|---|---|
| FC-sup | 91.9 | 77.1 | 14.8 | 71.5 | 64.0 | 7.5 |
| + ART | 91.8 | 78.7 | **13.1** | 69.3 | 64.8 | **4.5** |
| VIB | 87.8 | 82.6 | 4.2 | 65.4 | 63.6 | 0.7 |
| + ART | 87.6 | 87.0 | **0.6** | 65.8 | 65.5 | **0.3** |
| VIB-sup | 90.2 | 81.4 | 8.8 | 68.7 | 63.7 | 5.0 |
| + ART | 90.0 | 86.1 | **3.9** | 70.3 | 65.7 | **4.6** |

Table 6: Augmented Rationale Training (ART) reduces the effect of adversarial attacks. Ori: original input, Att: input with attack text.

## 7 Augmented Rationale Training

From our previous analysis on the trade-off between GR and AR (§6.2), it is clear that avoiding attack sentences in rationales is a viable way to make such models more robust. Note that this is not obvious by construction since the addition of attacks affects other parameters such as position of the original text and discourse structure, which may throw off the 'predictor' component of the model. As a more explicit way of encouraging 'rationalizers' to ignore spurious text, we propose a simple method called *augmented rationale training* (ART). Specifically, we sample two sentences at random from the Wikitext-103 dataset (Merity et al., 2017) and insert them into the input passage at random position, setting their pseudo rationale labels $z^{\text{pseudo}} = 1$ and the labels for all other sentences as $z = 0$. We limit the addition to only inserting two sentences to avoid exceeding the rationalizer maximal token limit. We then add an auxiliary negative binary cross entropy loss to train the model to *not* predict the pseudo rationale. This encourages the model to ignore spurious text that is unrelated to the task. Note that this procedure is both model-agnostic and does not require prior knowledge of the type of AddText attack.

Table 6 shows that ART improves robustness across the board for all models (FC-sup, VIB and VIB-sup) in both FEVER and MultiRC, dropping Δ scores by as much as 5.9% (VIB-sup on FEVER). We further analyzed these results to break down performance in terms of attack and gold sentence capture rate. Table 7 shows that ART greatly improves the percentage of sentences under the "Gold ✓ Attack ✗" category (31.8% → 65.4% for VIB and 11.3% → 63.5% for VIB-sup). This corroborates our expectations for ART and shows its effec-

tiveness at keeping GR high while lowering AR.

Interestingly, the random Wikipedia sentences we added in ART are not topically or contextually related to the original instance text at all, yet they seem to help the trained model ignore adversarially constructed text that is tailored for specific test instances. This points to the promise of ART in future work, where perhaps more complex generation schemes or use of attack information could provide even better robustness.

| | | | +ART | |
|---|---|---|---|---|
| | VIB | VIB-sup | VIB | VIB-sup |
| G ✓ A ✓ | 34.3% | 76.7% | 6.0% | 25.4% |
| G ✓ A ✗ | 31.8% | 11.3% | 65.4% | 63.5% |
| G ✗ A ✓ | 22.0% | 11.5% | 3.2% | 4.2% |
| G ✗ A ✗ | 12.0% | 0.4% | 25.4% | 6.8% |

Table 7: The percentage of examples in the development set (in four categories the same way as Table 5) of the VIB and VIB-sup models without (left) and with (right) ART training on the FEVER dataset.

## 8 Discussion

In this work, we investigated whether neural rationale models are robust to adversarial attacks. We constructed a variety of AddText attacks across five different tasks and evaluated several state-of-the-art rationale models. Our findings raise two key messages for future research in both interpretability and robustness of NLP models:

**Interpretability** We identify an opportunity to use adversarial attacks as a means to *evaluate* rationale models (especially extractive ones). In contrast to existing metrics like IOU used in prior work (DeYoung et al., 2020; Paranjape et al., 2020), robustness more accurately tests how crucial the predicted rationale is to the model's decision making. Further, our analysis reveals that even state-of-the-art rationale models may not be consistent in focusing on the most relevant parts of the input, despite performing well on tasks they are trained on. This points to the need for better model architectures and training algorithms to better align rationale models with human judgements.

**Robustness** For adversarial attack research, we show that extractive rationale models are promising for improving robustness, while being sensitive to factors like the attack position or word choices in the attack text. Research that proposes new attacks can use rationale models as baselines to assess their effectiveness. Finally, the effectiveness of ART points to the potential for data augmentation in improving robustness of NLP systems, even against other types of attacks beyond AddText.

We hope our results can inspire more research at the intersection of interpretability and robustness.

## Acknowledgement

We thank the members of the Princeton NLP group and the anonymous reviewers for their valuable comments and feedback. HC is supported by the Princeton Upton Fellowship. This research is also supported by a Salesforce AI Research Grant.

## References

Alexander Alemi, Ian Fischer, Joshua Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–421.

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

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Association for Computational Linguistics (ACL)*, pages 2963–2977.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Association for Computational Linguistics (ACL)*.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning (ICML)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen F. Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized nlp models. In *Association for Computational Linguistics (ACL)*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Association for Computational Linguistics (ACL)*, pages 31–36.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Nuno Miguel Guerreiro and André F. T. Martins. 2021. SPECTRA: Sparse structured text rationalization. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *North American Association for Computational Linguistics (NAACL)*, pages 1875–1885.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Association for Computational Linguistics (ACL)*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *North American Association for Computational Linguistics (NAACL)*, pages 3543–3556.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. In *Association for Computational Linguistics (ACL)*, pages 4459–4473.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (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 *North American Association for Computational Linguistics (NAACL)*.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *IEEE International Conference on Data Mining (ICDM)*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Vlad Niculae and F. T. André Martins. 2020. Lp-sparsemap: Differentiable relaxed optimization for sparse structured prediction. In *International Conference on Machine Learning (ICML)*.

Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. 2021. An empirical study on the relation between network interpretability and adversarial robustness. *SN Computer Science*, 2(1):1–13.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Association for Computational Linguistics (ACL)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Association for Computational Linguistics (ACL)*, pages 2931–2951.

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. 2021. Learning from the best: Rationalizing prediction by adversarial information calibration. In *Conference on Artificial Intelligence (AAAI)*.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Conference on Artificial Intelligence (AAAI)*.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Association for Computational Linguistics (ACL)*, pages 5609–5626.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 10314–10332.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4094–4103.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2020. Do feature attribution methods correctly attribute features? In *Conference on Artificial Intelligence (AAAI)*.

# A Appendix

## A.1 Model Details

**VIB details** The sentence or token level logits $s \in \mathbb{R}^L$ (A.2 describes how the logits are obtained) parameterize a relaxed Bernoulli distribution $p(z_t \mid x) = \text{RelaxedBernoulli}(s)$ (also known as the Gumbel distribution (Jang et al., 2017)), where $z_t \in \{0, 1\}$ is the binary mask for sentence $t$. The relaxed Bernoulli distribution also allows for sampling a soft mask $z_t^* = \sigma(\frac{\log s + g}{\tau}) \in (0, 1)$, where $g$ is the sampled Gumbel noise. The soft masks $z^* = (z_1^*, z_2^*, ..., z_T^*)$ are sampled independently to mask the input sentences such that the latent $z = m^* \odot x$ for training. The following objective is optimized:

$$\ell_{\text{VIB}}(x, y) = \mathop{\mathbb{E}}_{z \sim p(z|x;\theta)} \big[ -\log p(y \mid z \odot x; \phi) \big] + \beta \text{KL}\big[p(z \mid x; \theta) \mid\mid p(z)\big],$$

where $\phi$ denotes the parameters of the predictor $C$, $\theta$ denotes the parameters of the rationalizer $R$, $p(z)$ is a predefined prior distribution parameterized by a sparsity ratio $\pi$, and $\beta \in \mathbb{R}$ controls the strength of the regularization.

During inference, we take the rationale as $z_t = \mathbb{1}[s_t \in \text{top-}k(s)]$, where $s \in \mathbb{R}^L$ is the vector of token or sentence-level logits, and $k$ is determined by the sparsity $\pi$.

**VIB-sup details** With human raitonale supervision $z^*$, the objective below is optimized:

$$\ell_{\text{VIB-sup}}(x, y) = \mathop{\mathbb{E}}_{z \sim p(z|x;\theta)} \big[ -\log p(y \mid z \odot x; \phi) \big] + \beta \text{KL}\big[p(z \mid x; \theta) \mid\mid p(z)\big] + \gamma \sum_t -z_t^* \log p(z_t \mid x; \theta),$$

where $\beta, \gamma \in \mathbb{R}$ are hyperparameters. During inference, the rationale module generates the mask $z$ the same way as the VIB model by picking the top-$k$ scored positions as the final hard mask. The third loss term will encourage the model to predict human annotated rationales, which is the ability we expect a robust model should exhibit.

**SPECTRA details** SPECTRA optimizes the following objective:

$$\ell_{\text{SPECTRA}}(x, y) = -\log p(y \mid z \odot x; \phi),$$
$$z = \mathop{\text{argmax}}_{z' \in \{0,1\}^L} (\text{score}(z'; s; \theta) - \frac{1}{2} \left\| z' \right\|^2),$$

where $s \in \mathbb{R}^L$ is the logit vector of tokens or sentences, and a global $\text{score}(\cdot)$ function that incorporates all constraints in the predefined factor graph. The factors can specify different logical constraints on the discrete mask $z$, e.g a `BUDGET` factor that enforces the size of the rationale as $\sum_t z_t \leq B$. The entire computation is deterministic and allows for back-propagation through the LP-SparseMAP solver (Niculae and Martins, 2020). We use the `BUDGET` factor in the global scoring function. To control the sparsity at $\pi$ (e.g., $\pi = 0.4$ for 40% sparsity), we can choose $B = L \times \pi$.

**FC-sup details** The FC model can be extended to leverage human annotated rationales supervision during training (FC-sup). We add a linear layer on top of the sentence/token representation and obtain the logits $s \in \mathbb{R}^L$. The logits are passed through the sigmoid function into mask probabilities to optimize the following objective:

$$\ell_{\text{FC-sup}}(x, y) = -\log p(y \mid x; \phi) + \gamma \sum_t -z_t^* \log p(z_t \mid x; \phi, \xi),$$

where $z_t^*$ is the human rationale, $\xi$ accounts for the parameters of the extra linear layer, and the hyperparameter $\gamma$ is selected based on the original performance by tuning on the development set.

## A.2 Implementation Details

We use two `BERT-base-uncased` (Wolf et al., 2020) as the rationalizer and the predictor components for all the models and one BERT-base for the Full Context (FC) baseline. The rationales for FEVER, MultiRC, SQuAD are extracted at sentence level, and Beer and Hotel are at token-level.

$$\texttt{BERT}(x) = \big(\mathbf{h}_{\texttt{[CLS]}}, \mathbf{h}_0^1, \mathbf{h}_0^2, ..., \mathbf{h}_0^{n_0}, \mathbf{h}_{\texttt{[SEP]}}, \mathbf{h}_1^1, \mathbf{h}_1^2, ..., \mathbf{h}_1^{n_1}, ..., \mathbf{h}_T^1, \mathbf{h}_T^2, ..., \mathbf{h}_T^{n_T}, \mathbf{h}_{\texttt{[SEP]}}\big),$$

where the input text is formatted as *query* with sentence index 0 and *context* with sentence index 1 to $T$. For sentiment tasks, the 0-th sentence and the first `[SEP]` token are omitted. For sentence-level representations, we concatenate the start and end vectors of each sentence. For instance, the $t$-th sentence representation is $\mathbf{h}_t = [\mathbf{h}_t^0; \mathbf{h}_t^{n(t)}]$. For token-level representations, we use the hidden vectors directly. The representations are passed to a linear layer $\{\mathbf{w}, b\}$ to obtain logit for each sentence $s = \mathbf{w}^\intercal \mathbf{h}_t + b$.

**Training** Both the rationalizer and the predictor in the rationale models are initialized with pretrained BERT (Devlin et al., 2019). We predetermine rationale sparsity before fine-tuning based on the average rationale length in the development set following previous work (Paranjape et al., 2020; Guerreiro and Martins, 2021). We set $\pi = 0.4$ for FEVER, $\pi = 0.2$ for MultiRC, $\pi = 0.7$ for SQuAD, $\pi = 0.1$ for Beer, and $\pi = 0.15$ for Hotel. The hyperparameter $k$ (for top-$k$ ratioanle extraction) is selected based on the percentage $\pi$ of the human annotated rationales in the development set (following Paranjape et al. (2020)). During evaluation, for each passage $k = \pi \times$ #sentences. We select the model parameters based on the highest fine-tuned task performance on the development set. The models with rationale supervision will select the same amount of text as their no-supervision counterparts. The epoch/learning rate/batch size for the different datasets are described in Table A.2.

| Dataset | Epoch | Learing Rate | Batch Size |
|---|---|---|---|
| FEVER | 10 | 5e-5 | 32 |
| MultiRC | 10 | 5e-5 | 32 |
| SQuAD | 3 | 1e-5 | 32 |
| Beer | 20 | 5e-5 | 64 |
| Hotel | 20 | 5e-5 | 64 |

### A.3 Qualitative Examples

We provide qualitative examples of the rationale model predictions for each dataset in Table 8.

### A.4 Attack Position and Lexical Variation

Figure 4 shows a more fine-grained trend reflecting the sensitivity of AR against inserted attack position. As the attack position move from the beginning of the passage towards the end, AR decreases across all models. With ART training (R6 in §6), the AR also becomes less sensitive to positions. We also experimented with various adjectives related to appearance as the attack and observe the same trend. For example, when inserting "The carpet looks really ugly/beautiful." to the Beer dataset, VIB performance drops 93.8 → 83.1 while FC drops 93.8 → 61.6.

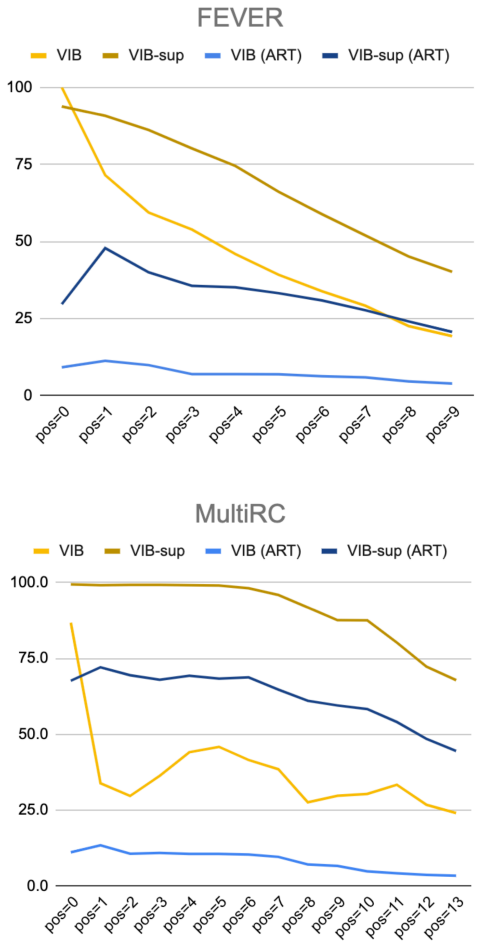


Figure 4: The attack capture rate (AR) changes with respect to different attack positions for FEVER and MultiRC.

| Dataset | Query | Passage | Predicted / Gold Label |
|---|---|---|---|
| FEVER | The Silver Surfer appears only in Icelandic comic books. | The Silver Surfer is a fictional superhero appearing in American comic books published by Marvel Comics. The character also appears in a number of movies , television , and video game adaptations. The character was created by Jack Kirby , and first appeared in the comic book Fantastic Four # 48 , published in 1966. The Silver Surfer is a humanoid with metallic skin who can travel space with the aid of his surfboard-like craft. Originally a young astronomer named Norrin Radd on the planet Zenn-La , he saved his homeworld from the planet devourer , Galactus , by serving as his herald. Imbued in return with a tiny portion of Galactus 's Power Cosmic , Radd acquired vast power , a new body and a surfboard-like craft on which he could travel faster than light. Now known as the Silver Surfer , Radd roamed the cosmos searching for planets for Galactus to consume. When his travels took him to Earth , he met the Fantastic Four , a team of powerful superheroes who helped him rediscover his humanity and nobility of spirit. Betraying Galactus , the Surfer saved Earth but was exiled there as punishment. The Carey Hayes appears only in scottish comic books. | Refutes / Refutes |
| MultiRC | What did Jenny and her friends enjoy when they walked to the sand ? \|\| The sun | Jenny was a 13 year old girl with blond hair and blue eyes . She had gotten out of her last day of school and was free for the summer. Two of her friends were going to the nearby beach to do some swimming and enjoy the sun. Jenny went with them and when they got there the beach was very full and there were people everywhere. They changed into their bathing suits and went to the water. The water was very cold. They chose not swim and walked to the sand. Then they laid down on some towels and enjoyed the sun. After several hours Jenny and her friends fell asleep. Jesse and her foe enjoy the moon when they walked to the sand. They woke up and the sun was beginning to set. When Jenny sat up she found that it was painful to touch her skin. When she looked down she saw that she had a very bad sunburn. Her friends were also very badly sunburned so they went home. Jenny 's mother gave her a cream to put on the sunburn. | False / True |
| SQuAD | When did oil finally returned to its bretton woods levels ? | This contributed to the "oil shock". After 1971, opec was slow to readjust prices to reflect this depreciation. From 1947 to 1967, the dollar price of oil had risen by less than two percent per year. Until the oil shock, the price had also remained fairly stable versus other currencies and commodities. Opec ministers had not developed institutional mechanisms to update prices in sync with changing market conditions, so their real incomes lagged. The substantial price increases of 1973 – 1974 largely returned their prices and corresponding incomes to bretton woods levels in terms of commodities such as gold. Oil finally returned to its colossus mickelson levels in 1898. | 1973-1974 / 1973-1974 |
| Beer | | The tea looks horrible. Poured from a 12oz bottle into a delirium tremens glass. This is so hard to find in columbus for some reason, but I was able to get it in toledo... murky yellow appearance with a very thin white head. The aroma is bready and a little sour. The flavor is really complex, with at least the following tastes: wheat, spicy hops, bread, bananas, and a toasty after - taste. It was really outstanding. I'd recommend this to anyone, go out and try it. I think it's the best so far from this brewery. | Positive / Positive |
| Hotel | | My car is very filthy. The hotel was in a brilliant location and very near a metro station. Yes the room was small but it was clean and very well equipped. the bathroom was a really good size and lets face it how long do you spend in your hotel room anyway? The breakfast was fantastic and the staff were really friendly and helpful. We will definately stay here when we return to barcelona. It's worth going up to the roof of the hotel for the view over the city. | Negative / Positive |

Table 8: Examples of predicted rationales (yellow highlight), gold rationale (cyan text), and attack (red text) for passages in different datasets.